CoDEl-BR: An Electoral Debate Corpus from Brazilian Municipal Elections

Alessandra Gomes[1], Aline Paes[1], Helena Caseli [2]
[1]Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brazil
[2]Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, Brazil
alegomes@id.uff.br, alinepaes@ic.uff.br, helenacaseli@ufscar.br

Abstract

Electoral debates are influential moments in public discourse, providing candidates with a high-visibility platform to present their proposals, contrast their positions, and engage in exchanges that shape voter decisions. In Brazil, these debates reach a broad and diverse audience, reflecting regional, social, and ideological variations that affect linguistic choices and thematic content. This paper presents CoDEl-BR (*Corpus de Debates Eleitorais*, in Portuguese), a corpus of transcripts from 22 second-round mayoral debates held in 13 Brazilian state capitals during the 2024 municipal elections. It comprises 2,943 transcript segments totaling approximately 32 hours. Exploratory analyses reveal differences in thematic priorities between candidates and voters' questions, as well as variations by race and party affiliation. The corpus aims to enable research in discourse and argumentation analysis, stance and sentiment detection, polarization modeling, and other related NLP tasks. We demonstrate that this initial release provides a curated, high-quality subset of debates with significant potential for expansion.

1 Introduction

Electoral debates play a crucial role in the democratic process, offering candidates a public platform to present their plans and articulate how they intend to implement them. For voters, they are a key source of information, enabling side-by-side comparisons as candidates exchange ideas, opinions, proposals, and solutions in response to questions from journalists, voters, and opponents.

In Brazil, election debates are broadcast on nationwide open television (Kantar IBOPE Media, 2025), providing candidates with a platform to deploy rhetorical strategies aimed at attracting voters. Brazil's vast territory and demographic complexity also bring together candidates from multiple parties, regions, genders, social classes, and racial backgrounds. This diversity reflects the country's multicultural and socio-political landscape, shaping the speeches' content, style, topics, interests, and concerns that emerge during the debates.

Given this linguistic and socio-political diversity, building a corpus of Brazilian Portuguese debate transcripts can provide a rich empirical resource for analyzing how candidates communicate, persuade, and position themselves across different themes. Such a corpus would hold the potential to support a wide range of research applications, including discourse analysis, misinformation detection, opinion and sentiment analysis, argumentation mining, topic modeling, and dialogue training, among others. It also enables interdisciplinary work at the intersection of Natural Language Processing (NLP), Political Communication, Digital Humanities, Computational Social Science and sociocultural studies on gender, race, and class.

To the best of our knowledge, there is no prior work that provides a corpus of electoral debate transcripts in Brazilian Portuguese. Existing resources in the political-electoral domain either focus on parliamentary debates (Braga and Brito, 2019; Cunha, 2017, 2025; Goffredo et al., 2022; dos Santos et al., 2023; Cochrane et al., 2022; Jiménez-Preciado et al., 2025; Tamper et al., 2022) or contain electoral debates in languages other than Portuguese (Fermín L. Cruz and Troyano, 2025; Mestre et al., 2021).

To fill this gap, this work presents CoDEl-BR, a corpus composed of transcripts from debates held during the second round of mayoral elections in Brazilian state capitals in the 2024 election. The elections for mayor in Brazil can be decided in the first or second round. Candidates win in the first round if they receive more than 50% of the valid votes. Otherwise, the two candidates with the most votes in the first round advance to the second round, and the candidate who obtains the majority of the votes in that final round wins the

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 647–656

election (de Bessa Dias, 2013). We selected the 2024 municipal elections because they represent the most recent nationwide electoral event. Among the two runs held in 2024 (mayor and councilor), we focused on the mayoral race, as it receives substantially greater media attention, including the publicly available debates.

This paper focuses on a well-defined and manageable subset of the data. We consider debates from the second round in capital cities, which provide a coherent and representative sample and establish a solid basis for future extensions. CoDEl-BR totals approximately 32 hours, compiling 2,943 transcripts of speeches among debate moderators, questioners, and 28 candidates of diverse gender, race, and party affiliation across 22 debates in 13 capitals. It displays a marked gender and racial imbalance, with most transcripts from male and white candidates, reflecting Brazil's political landscape. Key findings highlight differences in thematic priorities among candidates by race and party affiliation, as well as between candidates' speeches and the concerns raised by journalists and voters in their questions.[1]

This paper is organized as follows: Section 2 reviews related work; Section 3 describes the process of building CoDEl-BR; Section 4 provides an overview of CoDEl-BR, including its description and key statistics; and Section 5 concludes the paper and outlines directions for future work.

2 Related Works

Previous studies have introduced corpora or datasets of political debates, focusing predominantly on debates held in chambers or parliaments, where official transcripts are readily available. Examples include `ItaParlCorpus`, (Cova, 2025), `FREDSum`, (Rennard et al., 2023), `ASR Bundestag`, (Wirth and Peinl, 2024), and `PTPARL-D` (Almeida et al., 2021), comprising parliamentary debates from Italy, France, Germany, and Portugal, respectively. Other corpora originate from transcribed debates in the Finnish parliament (Simola et al., 2025) and an English corpus featuring debates among the permanent members of the United Nations Security Council (Efat et al., 2023).

There are also multilingual corpora and datasets that compile parliamentary debates across different countries. `ParlaSent-BCS` consists of transcripts from the parliaments of Bosnia-Herzegovina, Croatia, and Serbia (Mochtak et al., 2022). Similarly, Truan and Romary (2022) introduce a dataset that covers debates from the parliaments of the United Kingdom, Germany, and France. The `ParlaMint` project provides a multilingual collection comprising parliamentary debates from 29 European countries.[2]

The work most related to ours is (Fermín L. Cruz and Troyano, 2025), which introduces `DebatES`, a Spanish dataset comprising transcripts of televised debates from general elections in Spain between 1993 and 2023. In that work, the transcriptions were generated using `WhisperX`, presented in (Bain et al., 2023), rather than being obtained directly from official sources. Another related dataset, used in (Mestre et al., 2021), was created for a Kaggle competition[3] from YouTube videos of the 2020 US presidential debates and transcribed using `Rev`[4]. Although these datasets share similar goals with our work, none provide transcripts of election debates in Portuguese. Specifically for the electoral domain, we found no corpora or datasets containing transcripts of election debates in Portuguese.

3 The CoDEl-BR Construction

In the 2024 Brazilian elections, 15 of 26 state capitals held second-round mayoral contests[5]. Therefore, we built CoDEl-BR using transcripts derived from debate videos featuring the candidates who competed in these runoffs.

The construction of the debate corpus followed a seven-step multimodal pipeline, all implemented using the Python language and illustrated in Figure 1. The process was semi-automated: steps in red were performed algorithmically, while those in blue required human intervention. In steps one, two, and six, a human review was conducted to correct potential errors. Step five required manual audio annotation of some samples, and step seven involved prompt engineering to achieve the desired results.

[1] CoDEl-BR is available at https://github.com/alegomesbr/CoDEl-BR

[2] https://www.clarin.eu/parlamint#parlamint-corpora

[3] https://www.kaggle.com/datasets/headsortails/us-election-2020-presidential-debates

[4] https://www.rev.com/

[5] https://www.tse.jus.br/comunicacao/noticias/2024/Outubro/falta-1-dia-confira-quem-esta-na-disputa-no-2o-turno-das-eleicoes-2024

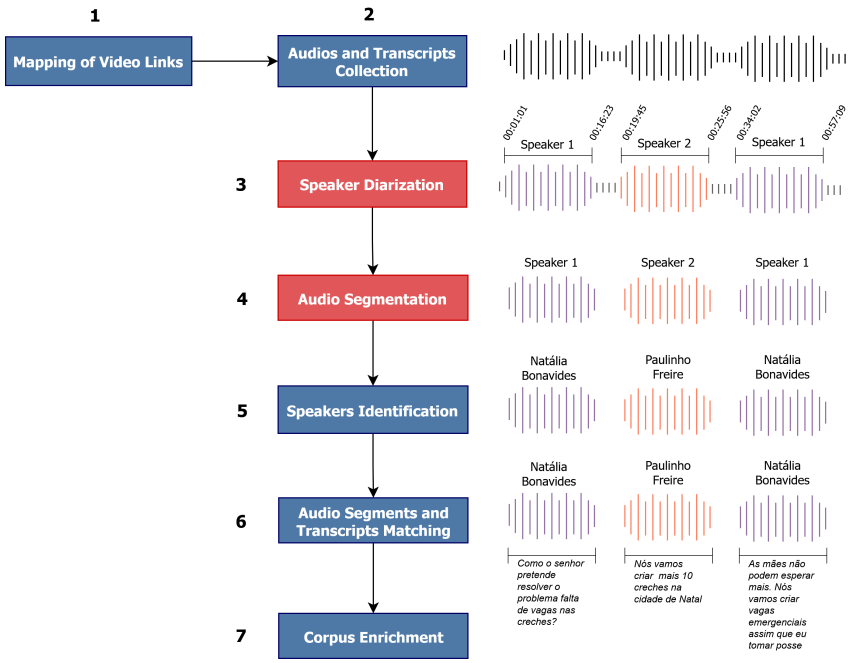

Figure 1: Overview of the seven-step multimodal pipeline for constructing CoDEl-BR

3.1 Video Links Mapping

In the first step, we mapped all YouTube video links of debates featuring candidates who advanced to the second round in the state capitals for the 2024 Brazilian municipal elections. To this end, we used the `YouTube Data API`[6] to collect links by combining the following keywords: "debate" "candidate's name on the ballot" and "city name" and "second round" and "2024 elections". The API returned videos that exactly matched the specified keywords and others containing related or approximated expressions. Consequently, a manual review was necessary to filter out links that did not correspond to second-round debates. At the end of this step, we mapped 26 valid video links from 13 cities: one for Belo Horizonte, Campo Grande, Cuiabá, Curitiba, Natal, and Palmas; two for Fortaleza, João Pessoa, Manaus, Porto Velho, and São Paulo; four for Porto Alegre, and five for Goiânia.

One debate from Goiânia was hosted on YouTube in four parts, but will be treated as a single video representing one debate. Thus, finishing the first step we had 22 valid videos to be processed. Besides, at the time of data collection, no debate videos were found for Belém or Aracajú. Although second-round debates took place in both cities, the broadcasting channels likely did not make videos available on YouTube.

[6] `https://developers.google.com/youtube/v3`

3.2 Audio and Transcripts Collection

Since our focus is on the spoken component of the videos rather than the visual content, the next step involved extracting the audio tracks and their corresponding transcripts. For audio extraction, we adopted the `YouTube Data API` in combination with the `YT-DLP` library[7]. Along with the audio extraction, we also collected the automatically generated Portuguese transcripts provided by YouTube. By the end of this step, we obtained the audio files encoded as `.flac` and the associated transcripts for all 22 videos.

3.3 Speaker Diarization

With the audio tracks and transcripts, we performed the speaker diarization (Ravanelli et al., 2024) to segment the audio by speaker. To perform this task, we relied on the pre-trained `speaker-diarization-3.19` pipeline from `pyannote`[8], an open-source toolkit for speaker diarization. The pipeline segments the audio into speeches, detects the number of speakers, generates embeddings for these speakers' voices and the speech segments, and applies clustering to assign each segment to its corresponding speaker.

The output consists of speaker labels along with the corresponding time interval of their speeches,

[7] `https://pypi.org/project/yt-dlp/`
[8] `https://huggingface.co/pyannote`

as illustrated in step three of Figure 1.

3.4 Audio Segmentation

In this step, we implemented a python script to segment each audio by splitting it according to the time intervals specified by the diarization process. Segmenting all 22 audios resulted in a total of 4,182 audio segments.

3.5 Speaker Identification

As illustrated in step 3 of Figure 1, diarization assigns only generic speaker labels. Therefore, it was necessary to identify the speakers and determine which segments corresponded to candidates or journalists. For this task, we used the pre-trained model `spkrec-ecapa-voxceleb`[9] from `SpeechBrain`[10], an open-source toolkit for Conversational Artificial Intelligence. This model performs speaker recognition by computing the cosine distance between voice embeddings (Ravanelli et al., 2024).

First, we manually annotated one representative audio segment for each of the 28 candidates. These served as ground truth for similarity matching with the remaining unannotated segments, adopting a threshold of 0.9. Audio segments that did not closely match any candidates were classified as belonging to journalists.

3.6 Audio-Transcripts Pairs Alignment

This step involves aligning the segmented and identified audio files with their respective transcripts. This requires segmenting the transcripts to precisely match each audio segment with the corresponding part of the text.

Using the timing metadata included in the collected transcriptions, we aligned them with the segmentation intervals and manually fixed any mismatches. This yielded 2,943 correctly aligned audio–transcript segments.

3.7 Corpus Enrichment

To enrich the corpus and expand its usability, we added candidates' gender, race, and party affiliation, as well as election results, sourced from Brazilian Superior Electoral Court Open Data Portal[11]. We also included an alternative transcript and topics identified for each segment.

Although the corpus already includes YouTube's automatic transcription, it contains errors and informal speech phenomena such as hesitations, self-corrections, and missing punctuation. Providing an additional corrected transcript enables tasks related to noise detection and correction.

We generated the additional transcriptions using `Whisper Timestamped`, an extension of `openai-whisper` that offers a multilingual Automatic Speech Recognition (ASR) with word-level timestamps and confidence scores (Louradour, 2023). Applying this ASR to each audio segment generated transcripts with writing corrections and a more formal style.

We employed LLMs to extract the main topics from each transcript segment, and to organize them into two attributes. The first one adopted a one-shot prompt that instructed the model to identify key topics using themes such as "Education", "Health", "Safety", "Economy", and others. This task yielded more than 800 distinct themes, leading us to run a second prompt. Then, we provided the LLMs with the themes from the first prompt and instructed them to group these into broader categories. The result was a group of 15 key topics categories: Corruption, Transparency and Ethics; Culture, Sport, Tourism and Leisure; Education; Environment and Sustainability; Health and Social Assistance; Infrastructure and Urban Development; Justice and Security; Partnerships and Relationships; Politics and Elections; Public Administration and Governance; Public Finance and Economics; Public Policies and Legislation; Public Services and Employment; Social Rights and Inclusion; Technology and Innovation. Segments with no topics previously detected were identified in a "No Topics Found" category.

Both tasks were carried out using three language model from different families: Mistral large (`mistral-large-3-25-12`)[12], Open AI `gpt-oss-20b`[13], and Llama 4 Scout (`llama-4-scout-17b-16e-instruct`)[14]. We selected only the intersection of key topics and key categories identified by all three models.

[9] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[10] https://huggingface.co/speechbrain

[11] https://dadosabertos.tse.jus.br/

[12] https://docs.mistral.ai/models/mistral-large-3-25-12

[13] https://openai.com/pt-BR/index/introducing-gpt-oss/

[14] https://ai.meta.com/blog/llama-4-multimodal-intelligence/

4 CoDEl-BR Overview

The CoDEl-BR comprises 2,943 transcripts segments (entries) from 142 unique speakers, counting about 32 hours and 318,085 words. Each entry corresponds to a full speech transcript, which can range from lengthy ones with multiple sentences to brief ones consisting of just a single sentence. This section presents a description of the CoDEl-BR corpus and provides a brief exploratory analysis, highlighting its key statistics.

4.1 Corpus Description

The corpus organizes data across 15 attributes, as presented in Table 1. Two are related to the debate: the identification of the speakers that participated in the debate, `speaker`, which can be a candidate (identified by name), journalist moderators (the professional leading the debate), journalist narrators (the professional explaining the debate rules), or questioner (journalists, specialists, or voters asking questions to candidates), and `city`, where the debate and election took place.

Four attributes are specifically related to the candidates: `party` – their party affiliation (AVANTE, MDB, PL, PMB, PODE, PP, PSD, PSOL, PT, and União), `gender` (Female or Male), `race` (White, Brown, or Black), and the candidate's outcome for the election, `election_result` (Elected or Not Elected). Attributes `gender` and `race` are associated only with candidates, as we only had access to the candidates' demographics information.

Finally, nine are related to the transcripts: `id_video_youtube`, the unique identifier for the YouTube video, `youtube_transcription`, the Portuguese transcription automatically generated by YouTube, `word_count_youtube`, the number of words in the YouTube transcription, `whisper_transcription`, the Portuguese transcription generated by OpenAI's Whisper model, `word_count_whisper`, number of words in the Whisper's transcription, `time_interval`, timestamp range within the transcripts segment, `duration`, length of the transcripts segment in seconds, `key_topics`, list of the key topics identified by the LLMs, and `key_topics_categories`, which groups these topics into at least one of 16 broad categories, including "No Topic Found".

4.2 Corpus Exploratory Analysis

This section presents an exploratory analysis of the corpus's key characteristics through two approaches: one focusing on examining speakers' demographic features and another on the main topics in their transcripts.

4.2.1 Demographics Analysis

CoDEl-BR has 142 different speakers, presented in Table 2: 28 candidates, 22 journalist moderators, 15 journalist narrators, and 77 questioners. Although most of the speakers are the questioners, the majority of the transcribed segments are from candidates and moderators, as expected from debate dynamics. Together, candidates' and moderators' transcripts account for 95.24% of the transcribed segments, 96.5% of the total duration, and 98.27% of the spoken words. The high average values for transcripts per debate, total speech time, and words per speech also highlight their prevalence.

The high number of questioners, combined with a low average of transcripts per debate, stems from the fact that most debates feature multiple questioners who each ask only one question. Moderators' long duration with low averages in duration and words per instance indicates their role as guides to the debate. Narrators generally showed low values overall but ranked second in total words, reflecting their limited contributions, which often consisted of lengthy speeches explaining the debate rules.

The high standard deviation values for candidates also reflect the dynamics of debates, where their speaking times vary from long speeches, sometimes up to three minutes, to short ones, as short as 30 seconds. Moderators exhibited a similar imbalance, though less extreme: they usually delivered long speeches at the beginning, end, and intervals, with brief interventions otherwise to simply conduct the debate. In contrast, narrators and questioners showed low and consistent standard deviation values, indicating their limited and uniform contributions during the debate.

Table 3 presents the demographic distribution of candidates, along with data on transcript segments, duration, and words. It reveals a strong gender and race imbalance: over 70% of segments belong to male and white candidates. By party affiliation, the majority of segments come from three parties (PL, PT, and União), represented by 16 candidates who account for over 50% (52.67%) of the total. Three other parties (AVANTE, PMB, and PSOL) each have one candidate, together comprising 10.71%. Election outcomes are evenly split between elected and non-elected candidates, as second-round results mean half secure victory and half face defeat.

Category	Attribute	Description
Debate	speaker	The speaker's identification.
	city	The city where the candidate ran for mayor.
Candidate	party	The speaker's party affiliation, if he or she is a candidate.
	gender	The speaker's gender, if he or she is a candidate.
	race	The speaker's race, if he or she is a candidate.
	election_result	The candidate's outcome for the election.
Transcript	id_video_youtube	The unique identifier for the YouTube video.
	youtube_transcription	The Portuguese transcription automatically generated by YouTube.
	word_count_youtube	The number of words in the YouTube transcription.
	whisper_transcription	The Portuguese transcription generated by OpenAI's Whisper model.
	word_count_whisper	The number of words in the Whisper's transcription.
	time_interval	The timestamp range within the transcription segment.
	duration	The length of the transcription segment in seconds
	key_topics	List of key topics identified in the transcript segment by the LLMs
	key_topics_categories	List of 16 broad categories for grouping the key topics, including "No Topic Found" category

Table 1: Attributes in CoDEl-BR and their descriptions

Speaker		Transcripts		Duration		Words	
Category	Total (%)	Total (%)	Avg (± Std)	Total (%)	Avg (± Std)	Total (%)	Avg (± Std)
candidate	28 (19.72%)	1,753 (59.56%)	62.61 (34.32)	1d:2h:41m (83.27%)	54s (42s)	274,944 (86.44%)	156.84 (117)
journalist moderator	22 (15.49%)	1,050 (35.68%)	42.04 (25.63)	4h:14m:25s (13.23%)	14s (30s)	37,628 (11.83%)	35.83 (59)
journalist narrator	15 (10.56%)	50 (1.70%)	3.33 (2.22)	26m:09s (1.36%)	31s (28)	3,339 (1.05%)	66.78 (64)
questioner	77 (54.23%)	90 (3.06%)	1.17 (0.41)	41m:05s (2.14%)	27s (15)	2,174 (0.68%)	43.48 (42)

Table 2: Distribution of segment transcriptions by speaker category in CoDEl-BR

Averages of segments per candidate reveal a wide range: the highest align with party affiliations, on MDB (93.5) and PP (100.50), while the lowest occur for race, with Black candidates (26). This occurs because MDB candidates Ricardo Nunes and Sebastião Salgado participated in 6 debates together, generating high number of transcribed segments. PP candidate Cícero Lucena inflated the count by dominating a free-format debate where his opponent spoke less and more slowly. Meanwhile, vice-candidates in Porto Velho had minimal speaking time in one debate, with one of them, a Black vice-candidate, delivering only 4 speeches and thus lowering the average.

The average speech duration shows little variation, as moderators controlled the timing in these debates. For words per segment, the most verbose were male candidates (160.80) and white candidates (163.35); by party, single candidates from PSOL (205.06) and AVANTE (173.39) topped the list, alongside MDB candidates (194.38). The even balance in election outcomes, combined with high standard deviation values, reflects the typical dynamics of electoral debates.

4.2.2 Key Topics Analysis

Figure 2 illustrates the distribution of key topic categories by speakers. Candidates and questioners agree on "Infrastructure and Urban Development" and "Health and Social Assistance" as top priorities, but diverge on others. Key mismatches include "Public Administration and Governance" and "Corruption, Transparency, and Ethics", highly important to candidates, while questioners rank

		Candidates	Transcriptions					
			Segments		Duration		Words	
Attribute	**Values**	**Total (%)**	**Total (%)**	**Avg (± Std)**	**Total (%)**	**Avg (± Std)**	**Total (%)**	**Avg (± Std)**
Gender	Male	20 (71.43%)	1,279 (72.96%)	63.95 (35.42)	19h:53m:53s (74.57%)	56s (43.44)	205,663 (74.80%)	160.80 (120.21)
	Female	8 (28.57%)	474 (27.04%)	59.25 (33.46)	6h:47m:7s (25.43%)	51s (40.35)	69,281 (25.20%)	146.16 (108.48)
Race	White	20 (71.43%)	1,247 (71.13%)	49.88 (33.41)	19h:30m:18s (73.10%)	56s (42.14)	203,705 (74.09%)	163.35 (118.48)
	Brown	6 (21.43%)	454 (25.90%)	41.27 (28.43)	6h:25m:52s (24.10%)	50s (44.30)	63,807 (23.21%)	140.54 (114.60)
	Black	2 (7.14%)	52 (2.97%)	26 (31.11)	44m:50s (2.80%)	51s (38.88)	7,432 (2.70%)	142.92 (97.47)
Party Affiliation	PL	8 (28.57%)	438 (24.98%)	54.75 (39.87)	6h:43m:18s (25.19%)	55.24 (43.40)	71,560 (26.03%)	163.37 (124.12)
	PT	4 (14.28%)	262 (14.94%)	65.50 (40.58)	4h:23m:49s (16.48%)	60.41 (42.00)	43,197 (15.71%)	164.87 (112.89)
	União	4 (14.28%)	225 (12.83%)	56.25 (12.44)	3h:44m:18s (14.01%)	59.81 (42.86)	38,511 (14.00%)	171.16 (117.45)
	PODE	3 (10.71%)	131 (7.47%)	43.67 (35.92)	1h:40m:46s (6.29%)	46.15 (30.81)	18,401 (6.69%)	140.46 (95.01)
	MDB	2 (7.14%)	187 (10.67%)	93.5 (37.47)	3h:26m:43s (12.91%)	66.32 (38.70)	36,350 (13.22%)	194.38 (109.03)
	PP	2 (7.14%)	201 (11.47%)	100.50 (62.93)	2h:17m:19s (8.58%)	40.99 (50.56)	19,868 (7.23%)	98.84 (116.62)
	PSD	2 (7.14%)	105 (5.99%)	52.5 (20.50)	1h:14m:28s (4.65%)	42.55 (36.59)	13,495 (4.91%)	128.52 (104.48)
	AVANTE	1 (3.58%)	61 (3.48%)	61 -	1h:03m:13s (3.95%)	62.18 (35.22)	10,577 (3.85%)	173.39 (97.69)
	PMB	1 (3.58%)	71 (4.05%)	71 -	40m:59s (2.56%)	34.63 (27.25)	8,220 (2.99%)	115.77 (84.31)
	PSOL	1 (3.58%)	72 (4.11%)	72 -	1h:26m:07s (5.38%)	71.76 (47.35)	14,765 (5.37%)	205.06 (130.22)
Election Result	Elected	14 (50.00%)	876 (49.97%)	35.04 (23.84)	13h:16m:58s (49.78%)	54s (42.18)	134,723 (49.00%)	153.79 (114.59)
	Not Elected	14 (50.00%)	877 (50.03%)	35.08 (23.72)	13h:24m:02s (50.22%)	55s (43.16)	140,221 (51.00%)	159.88 (119.94)

Table 3: Attributes of the candidates in CoDEl-BR

"Environment and Sustainability" third most important and "Security and Justice" on a par with "Health and Social Assistance."

An examination of the most frequent key topics among questioners and candidates, illustrated in Figure 3, reveals substantial overlap, but with differing priorities. Questioners' top 10 were: Urban Infrastructure, Security, Education, Health, Public Management, Public Transportation, Economic Development, Urban Planning, Environment, and Urban Mobility. Candidates' top 10: Public Management, Health, Education, Security, Corruption and Transparency, Urban Infrastructure, Public Transportation, General Infrastructure, Social Assistance, and Housing.

As for journalist moderators and narrators, they had no assigned topics for the majority of their key categories, which aligns with their roles in the debate.

Examining the distribution by gender, race, party affiliation, and election result in Figure 4 reveals that "Infrastructure and Urban Development" dominates across nearly all groups ($\sim$17% to 27%), followed by "Health and Social Assistance" ($\sim$11% to 20%), and "Education" ($\sim$5% to 16%). Less emphasized topics include "Technology and Innovation" ($\sim$3%), "Policies and Legislation" ($< 1\%$) and "Partnerships and Relations" ($< 2\%$).

When examining by attributes, the categories do not vary considerably by gender or election result,

	Candidate	Journalist Moderator	Journalist Narrator	Questioners
Infrastructure and Urban Development	18.14%	2.79%	0.0%	25.21%
Health and Social Assistance	16.71%	1.96%	0.0%	11.77%
Education	12.40%	1.40%	0.0%	8.40%
Public Administration and Governance	9.70%	0.56%	0.0%	5.89%
Security and Justice	9.14%	1.21%	0.0%	11.77%
Corruption, Transparency, and Ethics	7.70%	0.56%	0.0%	2.52%
Public Finances and Economy	7.25%	0.74%	0.0%	8.40%
Environment and Sustainability	3.05%	0.37%	0.0%	9.24%
Politics and Elections	2.84%	0.84%	2.00%	1.68%
Culture, Sports, Tourism, and Leisure	1.72%	0.0%	0.0%	2.52%
Public Services and Employment	1.36%	0.37%	0.0%	2.52%
Social Rights and Inclusion	0.95%	0.0%	0.0%	1.68%
Policies and Legislation	0.42%	0.0%	0.0%	0.84%
Partnerships and Relations	0.18%	0.0%	0.0%	0.0%
Technology and Innovation	0.70%	0.0%	0.0%	0.84%
No Topic Found	7.74%	89.20%	98.00%	6.72%

Figure 2: Distribution of the key topics categories by speaker in CoDEl-BR

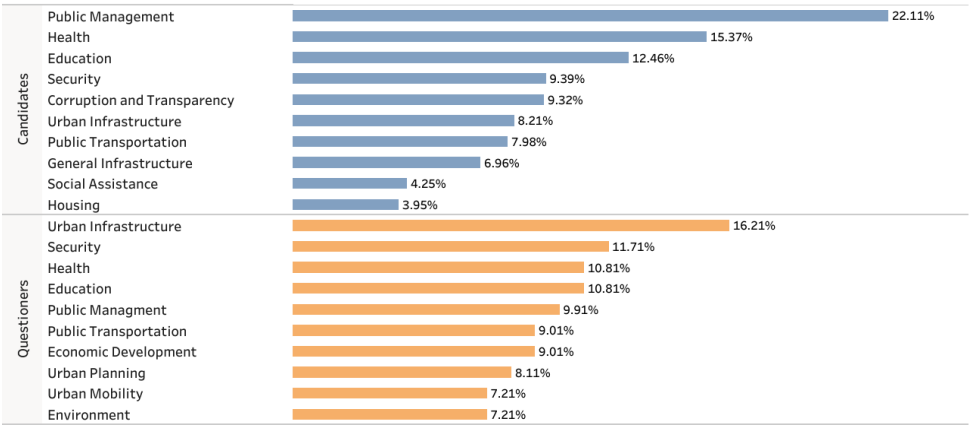

Figure 3: Top 10 key topics of candidates and questioners in CoDEl-BR

		Infrastructure and Urban Development	Health and Social Assistance	Education	Security and Justice	Public Administration and Governance	Corruption, Transparency, and Ethics	Public Finances and Economy	Environment and Sustainability	Politics and Elections	Culture, Sports, Tourism, and Leisure	Public Services and Employment	Social Rights and Inclusion	Technology and Innovation	Policies and Legislation	Partnerships and Relations	No Topic Found
GENDER	MALE	17.13%	16.09%	11.74%	10.26%	9.74%	7.59%	7.83%	3.15%	2.96%	2.00%	1.38%	1.05%	0.81%	0.53%	0.1%	7.64%
	FEMALE	20.97%	18.47%	14.25%	6.07%	9.63%	8.05%	5.67%	2.77%	2.51%	0.92%	1.32%	0.66%	0.4%	0.13%	0.4%	7.78%
RACE	WHITE	16.11%	15.96%	12.75%	10.21%	10.69%	8.77%	6.71%	3.11%	3.26%	1.73%	1.25%	1.1%	0.58%	0.53%	0.19%	7.05%
	BROWN	23.34%	18.73%	10.95%	6.49%	7.21%	4.32%	9.22%	3.17%	1.15%	1.87%	1.59%	0.58%	1.15%	0.14%	0.14%	9.95%
	BLACK	27.39%	19.18%	16.44%	4.11%	5.48%	9.59%	4.11%	0.0%	6.85%	0.0%	2.74%	0.0%	0.0%	0.0%	0.0%	4.11%
PARTY AFFILIATION	AVANTE	25.47%	13.21%	13.21%	13.21%	1.89%	5.66%	10.38%	1.89%	0.94%	3.77%	2.83%	0.94%	0.0%	0.0%	0.0%	6.6%
	MDB	25.31%	16.14%	6.96%	9.49%	10.76%	6.01%	13.29%	5.38%	0.63%	0.32%	1.27%	0.32%	0.95%	0.32%	0.32%	2.53%
	PL	15.11%	16.31%	12.97%	12.83%	9.36%	9.49%	6.42%	3.21%	4.28%	1.74%	0.8%	1.07%	0.66%	0.4%	0.0%	5.35%
	PMB	14.42%	11.54%	11.54%	6.73%	16.35%	12.5%	2.88%	1.92%	5.77%	0.0%	0.97%	1.92%	0.0%	0.0%	1.92%	11.54%
	PODE	15.6%	14.22%	15.14%	7.34%	13.3%	9.18%	6.88%	0.0%	4.13%	1.83%	2.29%	2.29%	0.46%	0.92%	0.46%	5.96%
	PP	18.38%	18.38%	14.34%	3.68%	4.78%	4.04%	5.88%	3.31%	0.73%	2.21%	1.84%	0.37%	2.57%	0.74%	0.37%	18.38%
	PSD	18.35%	12.03%	13.93%	4.43%	14.56%	6.96%	8.86%	6.96%	1.27%	0.0%	0.0%	0.0%	0.63%	0.63%	0.0%	11.39%
	PSOL	20.34%	14.41%	5.93%	16.95%	8.47%	15.25%	4.24%	0.0%	4.24%	0.0%	0.85%	1.69%	0.85%	0.0%	0.0%	6.78%
	PT	16.60%	19.65%	14.19%	8.73%	7.86%	7.64%	5.46%	3.49%	2.18%	2.62%	1.97%	0.87%	0.44%	0.44%	0.0%	7.86%
	UNIÃO	19.71%	20.00%	12.11%	5.92%	12.11%	4.51%	7.88%	1.69%	3.38%	2.54%	1.41%	0.85%	0.0%	0.28%	0.0%	7.61%
ELECTION RESULT	ELECTED	18.78%	15.99%	11.97%	7.74%	9.61%	6.59%	8.46%	3.3%	2.58%	2.08%	1.72%	1.00%	0.93%	0.57%	0.22%	8.46%
	NOT ELECTED	17.56%	17.42%	12.83%	10.49%	9.81%	8.78%	6.1%	2.81%	3.09%	1.37%	1.03%	0.89%	0.48%	0.27%	0.14%	6.93%

Figure 4: Distribution of the key topics categories according to Speaker's data descriptions in CoDEl-BR

but slightly by race. Notable highlights include black candidates discussing more about "Infrastructure and Urban Development", "Health and Social Assistance", "Education", and "Politics and Elections", brown candidates focusing on "Public Finances and Economy", and white candidates emphasizing "Security and Justice", and "Public Administration and Governance". Another key point is that black candidates did not address many categories such as "Environment and Sustainability" and "Culture, Sports, Tourism, and Leisure".

The major differences appear by party affiliation. AVANTE and MDB candidates exhibit high concentrations of speeches on "Infrastructure and Urban Development", with MDB also leading on "Public Finances and Economy". PMB candidates mainly focus on "Public Administration and Governance", PSD highlights on "Environment and Sustainability", PSOL discusses "Corruption, Transparency, and Ethics" and "Security and Justice" the most, and União on "Health and Social Assistance". In addition, MDB was the only party whose candidates addressed all key categories, indicating more diverse speeches, while PSD and PSOL candidates delivered more concentrated speeches and skipped four key categories each.

Another highlight is that PMB, PP, and PSD candidates all showed high percentages on "No Topic Found". This occurred mainly because their speeches in this category involved defending or attacking others or questioning the journalist moderator about the debate rules.

5 Conclusion and Future Works

This paper introduced CoDEl-BR, a new corpus of Brazilian Portuguese electoral debate transcripts that fills a gap in prior work, which has mainly focused on parliamentary contexts or other languages. The current release comprises transcripts from 22 second-round mayoral election debates held across 13 Brazilian state capitals in 2024. It was compiled through a semi-automated multimodal pipeline that involves video link mapping, audio and transcript collection, speaker diarization, speaker identification, and audio-transcript alignment. It is further enriched with topic annotations, a second automatic transcription, and candidates' metadata. Overall, CoDEl-BR comprises approximately 32 hours of audio and 2,943 speeches. CoDEl-BR exhibits a pronounced gender and racial imbalance, with over 70% of segments attributed to male and white candidates, reflecting structural inequalities in Brazilian politics.

Exploratory analyses reveal systematic differences in thematic priorities: candidates emphasize "Public Administration and Governance" and "Corruption, Transparency, and Ethics", while questioners highlighted "Environment and Sustainability". We also observed notable variations associated with race and party affiliation. For instance, Black candidates were more likely than their counterparts to address "Infrastructure and Urban Development", "Health and Social Assistance", "Education", and "Politics and Elections". On parties, MDB was the only one whose candidates covered all key categories, whereas PSD and PSOL candidates each skipped four.

CoDEl-BR establishes a foundational benchmark for Portuguese Political NLP, enabling advancements in topic modeling, stance detection, and discourse analysis for Brazilian electoral contexts. Future work will extend the corpus to additional cities and first-round debates, integrate textual and linguistic complexity metrics, and explore their relationship with electoral outcomes. We also plan to add annotations of persuasive strategies to support new NLP tasks on political persuasion.

Acknowledgments

This research was partially financially supported by CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*), grant 307088/2023-5, FAPERJ (*Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*), processes SEI-260003/002930/2024 and SEI-260003/000614/2023, CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) - Financial Code 001. We also thank the support of CNPq National Institutes of Science and Technology (INCTs), IAIA (grant 406417/2022-9), TILD-IAR (grant 408490/2024-1) and IAPROBEM (grant 408589/2024-8). This work is also in line with the UFSCar's Cátedra Computação Inteligente Centrada no Humano.

References

Paulo Almeida, Manuel Marques-Pita, and Joana Gonçalves-Sá. 2021. Ptparl-d: an annotated corpus of forty-four years of portuguese parliamentary debates. *Corpora*, 16(3):337–348.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech tran-

scription of long-form audio. *INTERSPEECH 2023*, pages 4489–4493.

Paloma Bernardino Braga and Daniel Martins de Brito. 2019. As funções figurativas do comentário metadiscursivo em debates eleitorais. *Cadernos de Linguagem e Sociedade*, 20(2).

Christopher Cochrane, Ludovic Rheault, Jean-François Godbout, Tanya Whyte, Michael W.-C. Wong, and Sophie Borwein. 2022. The automatic analysis of emotion in political speech based on transcripts. *Political Communication*, 39(1):98–121.

Joshua Cova. 2025. A new database for italian parliamentary speeches: introducing the itaparlcorpus dataset. *Italian Political Science Review*, 55(1).

Gustavo Ximenes Cunha. 2017. O papel dos conectores na co-construção de imagens identitárias: O uso do mas em debates eleitorais. *Alfa: Revista de Linguística*, 61(3):599–623.

Gustavo Ximenes Cunha. 2025. Conectores e seu encadeamento em inferências: um estudo do conector reformulativo "na verdade". *Confluência*, 69:109–145.

Renata Livia Arruda de Bessa Dias. 2013. A legitimidade das eleições majoritárias no brasil.

Matheus A. dos Santos, Nazareno Andrade, and Fabio Morais. 2023. Topic modeling of discussions in the standing committees of the brazilian chamber of deputies. *Journal of Information and Data Management*, 13(6).

Azher Ahmed Efat, Asif Atiq, Abrar Shahriar Abeed, Armanul Momin, and Md Golam Rabiul Alam. 2023. Empoliticon: Nlp and ml based approach for context and emotion classification of political speeches from transcripts. *IEEE Access*, 11:54808–54821.

F. Javier Ortega Fermín L. Cruz, Fernando Enríquez and José A. Troyano. 2025. Annotated spanish general election debate transcriptions 1993-2023. *Scientific Data*, 12:1578.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ana Lorena Jiménez-Preciado, José Álvarez García, Salvador Cruz-Aké, and Francisco Venegas-Martínez. 2025. The power of words from the 2024 united states presidential debates: A natural language processing approach. *Information*, 16(1).

Kantar IBOPE Media. 2025. Individual consumption within households – october/2025. https://kantaribopemedia.com/audiencia-de-video/. Acesso em 27 Novembro 2025.

Jérôme Louradour. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of The 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2022. The ParlaSent-BCS dataset of sentiment-annotated parliamentary debates from Bosnia-Herzegovina, Croatia, and Serbia. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, pages 132–139, Ljubljana, Slovenia.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. FREDSum: A dialogue summarization corpus for French political debates. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253, Singapore. Association for Computational Linguistics.

Salla Simola, Jeremias Nieminen, and Janne Tukiainen. 2025. Finnish parliamentary speeches dataset. *Scientific Data*, 12:1063.

Minna Tamper, Rafael Leal, Laura Sinikallio, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. 2022. Extracting knowledge from parliamentary debates for studying political culture and language. In *Proceedings of the International Workshop on Knowledge Graph Generation from Text (Text2KG 2022)*, pages 1–10, Crete, Hersonissos, Greece.

Naomi Truan and Laurent Romary. 2022. Building, encoding, and annotating a corpus of parliamentary debates in tei xml: A cross-linguistic account. *Journal of the Text Encoding Initiative*, 14. Selected Papers from the 2019 TEI Conference.

Johannes Wirth and René Peinl. 2024. Asr bundestag: A large-scale political debate dataset in german. In *Intelligent Systems and Applications*, pages 190–202, Cham. Springer Nature Switzerland.