Caminhos Discursivos Geométricos e a Geometria do Discurso

Alexandre Menezes Barroso

https://doi.org/10.1590/SciELOPreprints.15726

Submetido em: 2026-04-03
Postado em: 2026-05-05 (versão 1)
(AAAA-MM-DD)

Caminhos Discursivos Geométricos e a Geometria do Discurso

Alexandre Menezes Barroso
Universidade Estadual de Campinas
abarroso@outlook.com.br
https://orcid.org/0009-0005-2794-767X

Resumo

Este artigo propõe uma métrica de análise textual denominada *caminhos discursivos geométricos*: sequências de palavras intermediárias que conectam dois conceitos-âncora em um espaço vetorial e são filtradas por um critério de persistência dimensional. A análise compara, em espaços treinados separadamente, o corpus dos debates eleitorais para a prefeitura de São Paulo em 2024, tomado como discurso institucional, e um corpus de comentários online de referência, utilizado como interlocutor contrastivo. O estudo concentra-se em cinco pares de conceitos-âncora politicamente relevantes, interpretados qualitativamente com apoio de evidência sintética de recorrência e de checagens de robustez resumidas no artigo e detalhadas em material suplementar. Os resultados indicam, nos pares analisados, um contraste recorrente: no corpus do debate, as transições semânticas tendem a ser mediadas por intermediários técnico-administrativos ou institucionais; no corpus online de referência, tendem a atravessar atores políticos, eventos, antagonismos sociais e cadeias da economia cotidiana. A contribuição do artigo é dupla: metodologicamente, formaliza o conceito de "caminhos discursivos geométricos"; analiticamente, mostra como esse conceito pode ser incorporado à Linguística textual sem substituir a interpretação discursiva por técnicas computacionais.

Palavras-chave: análise textual; linguística textual; discurso político; debate eleitoral; semântica vetorial.

Geometric Discursive Paths and the Geometry of Discourse

Alexandre Menezes Barroso

Abstract: This article proposes a text-analytic metric called *geometric discursive paths*: sequences of intermediate words that connect two anchor concepts in a vector space and are filtered by a dimensional-persistence criterion. The analysis compares, in separately trained spaces, the corpus of the 2024 São Paulo mayoral debates, treated as institutional discourse, and an online comments reference corpus used as a contrastive interlocutor. The study focuses on five politically relevant anchor pairs, interpreted qualitatively with support from a compact layer of recurrence evidence and robustness checks summarized in the article and detailed in supplementary material. The results indicate, in the analyzed pairs, a recurrent contrast: in the debate corpus, semantic transitions tend to be mediated by technical-administrative or institutional intermediaries; in the online reference corpus, they tend to pass through political actors, events, social antagonisms, and everyday economic chains. The article makes a double contribution: methodologically, it formalizes the concept of "geometric discursive paths"; analytically, it shows how this concept can be incorporated into Textual Linguistics without replacing discourse interpretation with computational techniques.

Keywords: text analysis; text linguistics; political discourse; electoral debate; vector semantics.

1 Introdução

Em um ambiente de crescente polarização, os debates eleitorais para a prefeitura de São Paulo em 2024 condensaram tensões que ultrapassavam a política municipal. Desemprego, segurança, gestão urbana, desigualdade e relação entre esferas de poder apareceram como temas de campanha e de disputa pela definição dos problemas da cidade. Nesse cenário, o debate eleitoral interessa menos por *o que* diz do que por *como* organiza conexões entre conceitos politicamente densos.

Essa pergunta pode ser formulada em chave discursiva e se inscreve no debate recente sobre a reconfiguração digital da comunicação política, em que atuam "forças social e institucionalmente disruptivas" que "estão remodelando a comunicação política contemporânea" (ITUASSU et al., 2023, p. 7), em que as plataformas exercem "papel fundamental na circulação dessas ideias e na mobilização de pessoas" (OLIVEIRA; PILAU, 2025, p. 10) e em que podem circular "mensagens simples e focadas, direcionadas a segmentos específicos do eleitorado" (TUCCI; GOUVEIA, 2025, p. 18). Em termos bourdieusianos, a disputa eleitoral é também luta por capital simbólico e por formas legítimas de nomeação da realidade social; os sistemas simbólicos participam da imposição e legitimação de relações de força (BOURDIEU, 2016). Em Fairclough (2001), o discurso não apenas reflete relações de poder, mas ajuda a reproduzi-las ao moldar identidades, relações sociais e sistemas de conhecimento. Partimos daí para tratar o debate eleitoral como espaço em que conflitos são organizados e mediados.

O objetivo deste artigo é investigar como pares de conceitos politicamente relevantes são conectados por intermediários diferentes em dois espaços discursivos: de um lado, o corpus dos debates eleitorais paulistanos de 2024; de outro, um *corpus de comentários online de referência*, formado por comentários públicos de brasileiros no Reddit. O artigo não pretende representar "o brasileiro médio", nem equiparar esse corpus à população brasileira; importa reconhecer que, nas mídias digitais, "esses minipúblicos não se reúnem com base na lógica da deliberação sobre questões comuns [...] mas em torno do reforço identitário, da solidariedade afetiva e do fortalecimento de suas próprias visões e perspectivas" (ITUASSU et al., 2023, p. 19). Seu papel é servir de referência contrastiva para observar se, fora do debate institucional, operam "formas discursivas que não são estruturadas como a comunicação política típica" (OLIVEIRA, 2024, p. 16), se nelas é preciso ir além de "características linguísticas de nível superficial" (KRYKONIUK et al., 2024, p. 5, tradução nossa)[1], se podem emergir "demarcações de fronteiras entre o espectro político dos enunciantes e à atuação das instituições" (OLIVEIRA; PILAU, 2025, p.

[1] "surface-level linguistic features" (KRYKONIUK et al., 2024, p. 5)

20) e se há "disseminação ativa de informações distorcidas" (TUCCI; GOUVEIA, 2025, p. 21).

Para responder a essa pergunta, propomos um conceito: a métrica dos *caminhos discursivos geométricos*. Em síntese, ela identifica, em cada espaço vetorial, termos intermediários robustos entre conceitos-âncora, em diálogo com estudos sobre embeddings aplicados a "corpora em nível de sentença, como dados de mídias sociais" (CHEN et al., 2024, p. 2, tradução nossa)[2], com trabalhos em que tais abordagens "ainda produzem diferenças significativas entre as partes" (FREDÉN et al., 2024, p. 15, tradução nossa)[3] e, com *fine-tuning*, "produziu os resultados mais confiáveis e úteis" (FREDÉN et al., 2024, p. 17, tradução nossa)[4], além da ideia de espaço discursivo como "um sistema dinâmico de estados discursivos possíveis" (KRYKONIUK et al., 2024, p. 6, tradução nossa)[5]. O interesse do procedimento não está em substituir a interpretação discursiva por cálculo, mas em mostrar como corpora dispersos "produzem um discurso cuja coerência pode ser inferida" (OLIVEIRA, 2024, p. 17). A contribuição metodológica está no critério de persistência dimensional: um intermediário entra no caminho quando sua posição se mantém em múltiplas projeções, reduzindo aproximações fortuitas.

A contribuição do artigo é dupla. Primeiro, oferece um instrumento analítico novo para uma pergunta de Linguística textual: como descrever, de forma controlada, sequências intermediárias entre conceitos em um campo simbólico modelado vetorialmente, sem transformar a análise em demonstração de engenhosidade técnica. Segundo, usa esse instrumento para examinar cinco pares de conceitos fixados na análise — *democracia–ditadura*, *dinheiro–elite*, *governo–população*, *pobres–elite* e *população–desemprego* — e mostrar um contraste recorrente entre os dois corpora. No corpus do debate, as transições tendem a ser mediadas por termos ligados à administração, à institucionalidade e à legitimação técnico-burocrática; no corpus online de referência, tendem a aparecer intermediários mais ancorados em eventos, antagonistas nomeados e circuitos da vida econômica cotidiana.

Essa formulação exige prudência. O artigo não afirma que todo discurso institucional funcione desse modo, nem que o corpus online seja intrinsecamente "mais verdadeiro" ou mais próximo da experiência social. Tampouco sustenta que cinco pares de âncoras esgotem o espaço semântico dos corpora. Propõe-se aqui um estudo de caso contrastivo, orientado por pares focais e por uma métrica específica, capaz de oferecer evidência interpretável sobre distintas formas de mediação semântica. Por isso, o texto principal apresenta a conta metodológica necessária ao

[2] "sentence-level corpora like social media data" (CHEN et al., 2024, p. 2)

[3] "still produce meaningful differences between parties" (FREDÉN et al., 2024, p. 15)

[4] "produced the most reliable and useful results" (FREDÉN et al., 2024, p. 17)

[5] "a dynamic system of possible discourse states" (KRYKONIUK et al., 2024, p. 6)

argumento, enquanto o detalhamento auditável completo — especificação analítica, estatísticas de corpus e checagens adicionais — é remetido ao material suplementar.

A comparação deve ser entendida em seus próprios termos. O corpus dos debates e o corpus de comentários online de referência diferem em gênero, extensão e condições de produção. Essa assimetria é parte do desenho analítico: trata-se de comparar dois campos simbólicos constituídos em contextos distintos para perguntar se conectam os mesmos conceitos por mediações também distintas. O interesse do artigo não está em controlar todas as variáveis sociológicas possíveis, mas em observar diferença estrutural na organização das passagens semânticas.

O artigo organiza-se assim: a Seção 2 apresenta a fundamentação teórica; a 3, a metodologia; a 4, os corpora; a 5, os resultados; a 6, a discussão; e a 7, a conclusão.

2 Fundamentação teórica

Este é um trabalho de análise textual, e o seu ponto de partida teórico é a ideia de *texto como espaço relacional*. Não se trata de substituir a interpretação discursiva por cálculo, mas de construir uma heurística de leitura capaz de tornar observáveis certas mediações semânticas no interior do texto. A aposta do artigo é que relações de coocorrência lexical, quando modeladas em um espaço vetorial, podem oferecer indícios formais úteis para a análise textual, desde que permaneçam ancoradas em uma teoria do funcionamento do sentido e em uma interpretação discursiva historicamente situada.

Nesse quadro, a noção dos dois campos de Bühler – a *Zweifelderlehre* (1965, p. 119), delineada em *Sprachtheorie: Die Darstellungsfunktion der Sprache* (1965) – oferece um ponto de partida particularmente fértil. De acordo com o autor, a linguagem é compreendida como instrumento para falantes performarem atos de fala, *Sprechhandlungen* (BÜHLER, 1965, p. 53), orientados a interlocutores e a objetivos específicos. O contexto da fala estende-se, assim, por dois campos interconectados: (1) o *Symbolfeld*, composto pelo léxico e pelo conhecimento da língua; e (2) o *Zeigfeld*, isto é, o presente experiencial da produção dos enunciados (BÜHLER, 1990 apud HANKS, 2008, p. 151).

É relevante notar que Bühler define *Symbolfeld* como um ambiente sinsemântico, *synsemantische Umfeld* (BÜHLER, 1965, p. 81), no qual o sentido lexical se constitui enquanto parte de uma rede. A interpretação do texto depende, portanto, tanto de um quadro referencial (*Zeigfeld*) quanto do arranjo do léxico entre si, pela coocorrência de lexemas (*Symbolfeld*) (DIESSEL, 2012, p. 40). É precisamente nesse ponto que se insere a proposta deste trabalho:

investigar quantitativamente o campo simbólico, pela distribuição relacional das palavras, sem perder de vista que o valor interpretativo dessas relações só se estabiliza quando reconduzido ao campo demonstrativo, isto é, às condições sociais e históricas nas quais o discurso adquire sentido.

Partindo desse conceito de *Symbolfeld*, pensamos o texto de maneira sinsemântica. O texto, enquanto conjunto combinatório de lexemas, não é apenas uma sequência linear de unidades, mas um campo de relações. Se o compreendermos como espaço, as palavras ocupam posições relativas, definidas por aproximações, distâncias e oposições, em consonância com noções de campo influenciadas por Bourdieu (HANKS, 2008, p. 10). A ideia de *geometria discursiva*, tal como empregada aqui, nasce dessa intuição: sentidos não decorrem apenas de itens lexicais isolados, mas também da posição relacional que esses itens ocupam no interior de um campo simbólico.

Essa formulação, porém, deve ser entendida com prudência. A geometria discursiva, neste artigo, funciona como *heurística interpretativa ancorada em teoria textual*, e não como substituto da análise discursiva. O espaço vetorial não "explica" sozinho o texto; ele apenas torna visíveis regularidades relacionais que depois precisam ser interpretadas à luz do problema discursivo em questão. Em outras palavras, o cálculo organiza indícios; a análise lhes atribui pertinência.

O segundo movimento teórico diz respeito ao estatuto do debate eleitoral como unidade de análise. Consideraremos o debate eleitoral como um gênero discursivo em si mesmo, em sentido operacional. Podemos definir "gênero" como um conjunto relativamente estável de convenções e uma atividade socialmente aprovada com processos particulares de produção, distribuição e consumo (FAIRCLOUGH, 2001, p. 161). Essa definição é metodologicamente decisiva porque permite tratar a totalidade das falas de todos os candidatos como um único corpus. Os candidatos são adversários no processo político, porém, do ponto de vista do gênero, são colaboradores na construção do campo simbólico do debate: participam de uma mesma cena enunciativa, respondem às mesmas coerções formais e contribuem para a constituição de um mesmo objeto discursivo.

A formulação de Hanks (2008, p. 52) é útil aqui: o gênero, concebido como conjunto de elementos prototípicos utilizados de modos diferentes pelos autores sociais, não se fixa em uma estrutura unitária fechada; permanece parcial e transicional. Para os fins deste artigo, isso significa que não trataremos as falas de Nunes, Boulos, Marçal, Datena, Tabata ou Marina como espaços de texto autônomos. A unidade analítica não é o idioleto político individual, mas o debate eleitoral enquanto forma discursiva institucionalmente reconhecível. Essa decisão não

nega diferenças ideológicas entre candidatos; apenas as suspende provisoriamente para descrever o campo simbólico comum produzido por sua interação competitiva.

O terceiro pilar da seção é a hipótese da tecnocratização. Em Fairclough (2001), o discurso participa da naturalização de relações de poder também quando traduz problemas sociais e conflitos históricos em termos técnico-administrativos. Esse deslocamento não elimina o conflito, mas o reformula de maneira a privilegiar mediações burocráticas, gerenciais ou procedimentais. No quadro deste artigo, essa hipótese funciona como chave de leitura para os resultados: se os caminhos discursivos geométricos do corpus de debate tenderem a interpor mediadores institucionais entre conceitos-âncora, teremos um indício de tecnocratização discursiva no nível das transições semânticas.

É nesse ponto que a comparação com um corpus online de referência se torna teoricamente necessária. O corpus de comentários de internautas brasileiros não é mobilizado aqui como equivalente ao "povo" em sentido sociológico forte, nem como retrato representativo da totalidade da experiência social brasileira. Ele funciona, mais modestamente, como *interlocutor contrastivo* para o campo simbólico do debate. Sua utilidade analítica reside em oferecer um segundo espaço de texto, exterior ao debate institucional, contra o qual se pode observar se determinadas conexões semânticas são mediadas de modo diferente.

Com parcimônia, convém acrescentar que esse corpus de referência também permite vislumbrar a incidência de memória discursiva e de intertextualidade em circulação no comentário online. Como lembra Koch (2013), certos sentidos dependem da ativação de textos, eventos e referências já sedimentados na memória dos interlocutores. Quando, nos resultados, surgem intermediários ligados a personagens políticos, antagonismos sociais ou acontecimentos traumáticos, a interpretação pode ser lida como reativação dessa memória discursiva em um espaço de comentário mais permeável à nomeação explícita de conflitos.

Os três movimentos teóricos da seção convergem, assim, para a operacionalização apresentada a seguir. Bühler oferece a base para pensar o texto como campo relacional; a noção de gênero legitima o tratamento do debate como unidade analítica; e Fairclough fornece a chave para interpretar, no plano discursivo, a eventual mediação técnico-administrativa das conexões semânticas observadas.

3 Metodologia

Esta seção explicita o procedimento analítico no nível necessário para auditoria. Indicamos o que foi feito, as decisões interpretativas e os critérios de corpus, coerentes com a

"anotação linguística de corpora" (PEREIRA E SILVA, 2021, p. 2369), remetendo estatísticas, checagens e casos-limite ao suplemento. Essa economia expositiva é compatível com a construção de um "pipeline de validação seguro e confiável" (RUYTERS et al., 2025, p. 28, tradução nossa)[6], quando o artigo resume a arquitetura analítica e desloca detalhes operacionais ao apoio suplementar.

Vetorização do corpus

Cada corpus gera um espaço vetorial próprio. Em vez de projetar debate e comentários num único modelo, partimos de vetores fastText pré-treinados para o português e os ajustamos separadamente para cada corpus, solução coerente com cenários em que "usar sua metodologia de treinar diretamente em todos os dados disponíveis não era viável em nosso caso" (FREDÉN et al., 2024, p. 9, tradução nossa)[7]. A comparação, portanto, não se dá entre distâncias absolutas de um espaço único, mas entre caminhos obtidos para os mesmos pares de conceitos-âncora em dois espaços treinados de modo independente.

Em ambos os casos, os vetores têm 300 dimensões e partem do mesmo recurso inicial para o português. Mantêm-se constantes, nos dois modelos, o regime geral de treinamento, a amostragem negativa e a taxa de aprendizado. Preservamos, assim, um desenho interpretável para manter "transparência e explicabilidade" (RUYTERS et al., 2025, p. 30, tradução nossa)[8] nas comparações entre corpora treinados separadamente, sem depender de arquitetura opaca. O que varia entre os corpora é a configuração local do modelo de Word2Vec: no corpus dos debates, utilizou-se CBOW com janela 7 e frequência mínima 2; no corpus de comentários, utilizou-se Skip-gram com janela 10 e frequência mínima 8. Essa diferença não inviabiliza a comparação porque a unidade comparada não é uma medida absoluta entre vetores de corpora distintos, mas a forma das transições semânticas produzidas em cada espaço para os mesmos âncoras. Em vez de reduzir o contraste a um eixo único e "simplificar posições ideológicas em um espectro binário" (CHEN et al., 2024, p. 3, tradução nossa)[9], interessa-nos observar como cada corpus organiza mediações entre termos relevantes em um espaço que evita, tanto quanto possível, os "riscos potenciais trazidos por anotações incertas" (CHEN et al., 2024, p. 3, tradução nossa)[10].

Os gráficos bidimensionais têm função apenas ilustrativa. Os cálculos dos caminhos

[6] "secure and reliable validation pipeline" (RUYTERS et al., 2025, p. 28)

[7] "using their methodology of directly training on all the available data was not feasible in our case" (FREDÉN et al., 2024, p. 9)

[8] "transparency and explainability" (RUYTERS et al., 2025, p. 30)

[9] "simplify ideological positions into a binary spectrum" (CHEN et al., 2024, p. 3)

[10] "potential risks brought by uncertain annotation" (CHEN et al., 2024, p. 3)

discursivos geométricos são feitos no espaço completo de 300 dimensões, não nas reduções visuais, porque o objetivo metodológico não é parar em indícios locais de proximidade, mas "ir além de características linguísticas de nível superficial e engajar-se com as dimensões contextuais mais amplas da linguagem" (KRYKONIUK et al., 2024, p. 5, tradução nossa)[11].

Definição operacional de caminho discursivo

Chamamos de *conceitos-âncora* as palavras a partir das quais observamos uma transição semântica no espaço vetorial. Trabalhamos com cinco pares: *democracia–ditadura*, *dinheiro–elite*, *governo–população*, *pobres–elite* e *população–desemprego*. Em cada corpus, buscamos palavras que ocupem, de modo robusto, posições intermediárias entre os dois termos.

O procedimento segue quatro passos. Primeiro, centraliza-se o espaço vetorial em torno de sua média e calcula-se uma base de componentes principais sobre a totalidade das 300 dimensões. Segundo, percorrem-se projeções dimensionais decrescentes do espaço, do conjunto completo de dimensões até projeções reduzidas, para observar quais palavras permanecem como candidatas intermediárias sob perspectivas diversas do mesmo campo semântico. Em cada uma dessas projeções, toma-se o ponto médio entre os dois âncoras e selecionam-se os dez vocábulos mais próximos desse ponto. Entre eles, só são retidos como candidatos os itens cuja posição projetada efetivamente caia entre os dois âncoras; isso evita tratar como elo discursivo uma palavra próxima do ponto médio, mas semanticamente deslocada para fora do segmento que liga os conceitos.

Terceiro, acumula-se para cada candidato um escore de persistência dimensional. Uma palavra é considerada robusta quando reaparece como intermediária em mais de 30% das projeções examinadas. Esse limiar é conservador: um valor mais alto produziria caminhos mais curtos, mas seria menos adequado a corpora ainda modestos. O interesse é distinguir mediações ocasionais de padrões recorrentes e permitir um levantamento quantitativo dessas ocorrências.

Quarto, as palavras robustas são ordenadas segundo sua posição média projetada entre os dois âncoras em subconjuntos de dimensões mais altas. O resultado não é um encadeamento ganancioso de vizinhança passo a passo, mas uma sequência ordenada de intermediários persistentes entre dois polos semânticos. É esse critério de persistência dimensional, e não apenas a proximidade local entre vizinhos, que constitui a contribuição metodológica central do artigo. Em termos interpretativos, ele permite tratar o caminho discursivo como propriedade global do espaço, e não como efeito acidental de uma única projeção, aproximando-se da concepção de

[11] "move beyond surface-level linguistic features and engage with the broader contextual dimensions of language" (KRYKONIUK et al., 2024, p. 5)

que o espaço discursivo funciona como “um sistema dinâmico de estados discursivos possíveis” (KRYKONIUK et al., 2024, p. 6, tradução nossa)[12].

Especificação analítica resumida

Item	Especificação resumida
Recurso inicial	Vetores fastText pré-treinados para o português (Common Crawl).
Dimensionalidade	300 dimensões em ambos os corpora.
Treinamento do modelo	Ajuste do espaço com Word2Vec em modelos treinados separadamente para cada corpus.
Arquitetura	Debates: CBOW; comentários: Skip-gram.
Janela de contexto	Debates: 7; comentários: 10.
Frequência mínima	Debates: 2; comentários: 8.
Épocas	10.
Amostragem negativa	15.
Taxa de aprendizado	Inicial 0.025; mínima 0.0001.
Base de projeção	PCA calculada sobre o espaço centralizado; semente fixa para reprodutibilidade.
Projeções para persistência	Projeções dimensionais decrescentes de 300 até 2 dimensões.
Candidatos por projeção	Dez itens mais próximos do ponto médio entre os âncoras.
Critério de inclusão	Apenas candidatos cuja projeção caia entre os dois âncoras.
Regra de persistência	Manutenção de itens com ocorrência superior a 30% das projeções.
Ordenação final	Posição média projetada em subconjuntos de dimensões mais altas.
Pares-âncora do artigo	democracia–ditadura; dinheiro–elite; governo–população; pobres–elite; população–desemprego.
Detalhamento adicional	Estatísticas de corpus, pré-processamento, checagens de robustez e casos-limite no material suplementar.

Tabela 1: Especificação analítica resumida

A Tabela 1 reúne o mínimo necessário no corpo do artigo; o restante vai ao suplemento para preservar economia e auditabilidade. Mesmo em um cenário em que arquiteturas sentenciais já podem “superar a metodologia de nível de palavra de referência na medição de similaridades de posição em conjuntos de dados políticos do Twitter” (CHEN et al., 2024, p.

[12] “a dynamic system of possible discourse states” (KRYKONIUK et al., 2024, p. 6)

7, tradução nossa)[13], o ganho analítico aqui está em outra pergunta: como diferentes corpora organizam mediações semânticas robustas entre conceitos-âncora.

Robustez e escopo probatório

As interpretações desenvolvidas nas seções analíticas apoiam-se em cinco pares-âncora centrais, escolhidos por sua relevância política e por sua presença nos dois vocabulários. A força probatória do artigo é proporcional a esse desenho porque "a confiabilidade dos resultados pode variar mesmo dentro da mesma configuração de modelo, dependendo dos termos que são explorados" (FREDÉN et al., 2024, p. 16, tradução nossa)[14]. Também por isso, buscamos uma validação interpretável, interessada em "não apenas classificar os enunciados" (PEREIRA E SILVA, 2021, p. 12), mas em leituras que possam "coincidir com o senso comum social, as notícias políticas, bem como pesquisas anteriores" (CHEN et al., 2024, p. 7, tradução nossa)[15], sem pretender cartografar todo o discurso político brasileiro, e sim sustentar "a interpretação e a comparação de significado entre textos e registros" (KRYKONIUK et al., 2024, p. 12, tradução nossa)[16] a partir de trajetórias informativas.

Por isso, explicitamos três frentes mínimas de robustez: estabilidade das rotas em reexecuções, sensibilidade ao tamanho do corpus de comentários e comparação com versões menos restritivas do procedimento. Nessa chave, as checagens não entram como apêndice decorativo, mas como etapas "não como complementares, mas como necessárias" (RUYTERS et al., 2025, p. 45, tradução nossa)[17] para sustentar que os caminhos reportados não dependem de uma configuração contingente. No corpo do artigo, resumimos o necessário; no suplemento, apresentamos resultados e critérios de desempate ou exclusão para que outros possam "generalizar, complementar ou refutar esses dados" (PEREIRA E SILVA, 2021, p. 2393), em linha com a transparência e reprodutibilidade de outros estudos e pesquisas similares.

Metodologicamente, a tese deve ser lida em escala adequada. O que demonstramos é uma tendência contrastiva recorrente nos pares analisados e nas checagens de robustez reportadas, com a cautela adicional de que "os termos ideológicos apresentam menor estabili-

[13] "outperform benchmark word-level methodology on measuring position similarities in political Twitter datasets" (CHEN et al., 2024, p. 7)

[14] "the reliability of the results can vary even within the same model setting depending on the terms that are explored" (FREDÉN et al., 2024, p. 16)

[15] "coincide with social common sense, political news, as well as prior research" (CHEN et al., 2024, p. 7)

[16] "the interpretation and comparison of meaning across texts and registers" (KRYKONIUK et al., 2024, p. 12)

[17] "not as complementary, but as necessary" (RUYTERS et al., 2025, p. 45)

dade" (FREDÉN et al., 2024, p. 15, tradução nossa)[18], e não uma essência imutável do debate eleitoral ou do corpus online de referência. Essa delimitação fortalece o argumento: ao circunscrever o alcance probatório da métrica, tornamos mais nítido o que os caminhos discursivos geométricos efetivamente calculam e por que sua novidade reside em tornar observáveis, de forma comparável, mediações semânticas robustas entre conceitos politicamente relevantes.

4 Corpus

A análise compara dois espaços discursivos construídos a partir de corpora distintos em gênero, extensão e condições de produção. Por isso, esta seção tem uma função principalmente descritiva e defensiva: explicitar de onde vieram os dados, como foram consolidados e em que sentido a comparação é válida. O primeiro corpus corresponde ao debate eleitoral paulistano de 2024; o segundo, doravante denominado *corpus de comentários online de referência*, reúne comentários públicos coletados no Reddit. Quando empregamos adiante a expressão "discurso popular", ela funciona apenas como abreviação operacional para esse segundo conjunto, e não como designação sociológica exaustiva.

O corpus dos debates reúne 13 horas, 4 minutos e 13 segundos de material referente às eleições para a prefeitura de São Paulo em 2024, abrangendo os debates da Band (primeiro turno), Record (primeiro e segundo turnos), TV Cultura (primeiro turno), Terra (primeiro turno) e UOL (primeiro turno). A transcrição foi realizada automaticamente com modelo Whisper *medium* para português, em segmentos sucessivos de 60 segundos, posteriormente concatenados em ordem cronológica. Esse procedimento permitiu lidar de modo relativamente estável com um material marcado por múltiplos falantes, variações de volume, sobreposições e interrupções. Após a transcrição, o texto passou por pós-processamento: remoção de artefatos não linguísticos, normalização de pontuação e formatação, verificação de continuidade entre segmentos e correção manual de erros evidentes, sobretudo em nomes próprios e topônimos (por exemplo, correções do tipo "De Atena" para "Datena"). As transcrições foram então reunidas, sem indicação de falante, em um único arquivo de texto. Essa decisão decorre da unidade analítica adotada no artigo: interessa-nos o debate enquanto gênero discursivo, e não a modelagem separada de idioletos concorrentes.

O *corpus de comentários online de referência* reúne 148.792 comentários públicos de usuários brasileiros em fóruns do Reddit, coletados por meio da API oficial da plataforma. A estratégia de coleta buscou diversidade temática e de posicionamento discursivo, combinando

[18] "the ideological terms show less stability" (FREDÉN et al., 2024, p. 15)

comunidades voltadas a política, economia, cotidiano e temas locais. Foram incluídos os fóruns "Brasil", "BrasildoB", "Brasilivre", "Conversas", "Desabafos", "Investimentos" e "SaoPaulo". A lógica de amostragem partiu da seleção de tópicos por comunidade e da coleta integral das respectivas cadeias de comentários, de modo a preservar, tanto quanto possível, a estrutura conversacional do material. A distribuição do corpus por fórum aparece na Tabela 2.

Fórum	**Qtde. Comentários**
Brasil	34.034
BrasildoB	15.716
Brasilivre	22.767
Conversas	28.272
Desabafos	5.878
Investimentos	20.458
SaoPaulo	21.667
Total	148.792

Tabela 2: Quantidade de comentários por fórum no corpus de referência

Depois de coletado, o material textual foi agregado em um único arquivo e submetido a limpeza básica antes da modelagem: remoção de URLs, emojis, elementos próprios da plataforma, identificadores numéricos e outros resíduos não textuais, além de padronização de espaçamento, normalização de caracteres e uniformização em caixa baixa. Mantivemos essa limpeza em nível deliberadamente moderado, suficiente para a modelagem lexical, mas sem pretensão de eliminar toda a variação gráfica típica de comentários online. O detalhamento completo do pipeline encontra-se no material suplementar.

Do ponto de vista ético, a composição dos corpora restringiu-se a materiais textuais de acesso público: transcrições de debates eleitorais transmitidos publicamente e comentários online, i.e., materiais textuais de acesso público. No corpus online, não foram coletados nem armazenados nomes de usuário, identificadores de perfil, metadados pessoais ou quaisquer outros elementos que permitissem identificação individual. Mais importante para a interpretação do artigo, porém, é reconhecer o estatuto comparativo desse material. O corpus de comentários online de referência não é representativo do Brasil nem pretende funcionar como sinônimo de "sociedade brasileira". Trata-se de um interlocutor contrastivo para o discurso institucional do debate. Em consequência, a comparação aqui proposta não busca equiparar sociologicamente os corpora nem controlar todas as variáveis de gênero, extensão ou circulação. Seu objetivo é mais

circunscrito: observar se dois campos simbólicos construídos em contextos de produção distintos conectam os mesmos conceitos por mediações diferentes. O valor heurístico do segundo corpus está, portanto, menos em representar uma totalidade social do que em oferecer um termo de contraste relativamente heterogêneo fora do enquadramento institucional do debate. Por essa mesma razão, as checagens de sensibilidade ao tamanho do corpus online e outras verificações de robustez são resumidas no artigo e detalhadas no suplemento.

5 Resultados e análise

Os resultados apresentados a seguir correspondem aos cinco pares de conceitos-âncora efetivamente fixados na análise: "democracia"–"ditadura", "dinheiro"–"elite", "governo"–"população", "pobres"–"elite" e "população"–"desemprego". Esses pares foram escolhidos autoralmente por sua centralidade político-discursiva e por cobrirem domínios distintos do corpus. Eles são tratados aqui como *casos focais*: não esgotam o espaço semântico dos corpora, mas permitem observar, como os dois espaços vetoriais conectam conceitos de alta carga política por mediações diferentes. A força probatória da seção não depende apenas da leitura impressionista de cinco exemplos isolados. Ela se apoia também na recorrência das famílias de intermediários que reaparecem ao longo dos casos e na distinção entre itens fortemente interpretáveis e itens mais opacos. Checagens adicionais de estabilidade e sensibilidade são remetidas ao material suplementar; no corpo do artigo, privilegiamos a inteligibilidade do padrão e a disciplina interpretativa.

Tabela 3: Síntese dos cinco pares focais

Par de âncoras	Tipo predominante de intermediário no debate	Tipo predominante de intermediário no corpus online	Observação sintética
Democracia → Ditadura	mediação político-institucional e gestão urbana	memória histórica, eventos traumáticos, antagonismo político	o conflito é burocratizado no debate e historicizado no corpus online
Dinheiro → Elite	legitimação institucional e administrativa	operações econômicas cotidianas, acumulação e desigualdade	a elite é mediada por instâncias formais no debate e por experiência econômica no corpus online
Governo → População	inventário de gestão pública e controvérsias absorvidas administrativamente	poder estatal concreto, violência e territorialização	a população surge como destinatária de gestão no debate e como objeto de conflito no corpus online
Pobres → Elite	integração social burocratizada, com criminalização intermediária	antagonismo de classe e nomeação direta de dominantes	o debate dilui o conflito; o corpus online o nomeia
População → Desemprego	causalidade gerencial, especialistas e solução ideológica	indicadores sociais concretos e atores financeiros	o debate gerencializa o problema; o corpus online o estrutura socialmente

Antes dos casos, convém explicitar a camada agregada mínima de evidência que orienta sua leitura. Considerando apenas os intermediários mais interpretáveis dos cinco caminhos, o corpus do debate recorre repetidamente a três famílias léxico-discursivas: (i) instâncias institu-

cionais e burocráticas (“Ministério”, “Parlamentares”, “Administração”); (ii) vocabulário técnico-administrativo de gestão (“Gestão”, “Especialistas”, “Smart Sampa”, “UBSs”); e (iii) soluções politicamente marcadas apresentadas em chave gerencial (“Empreendedorismo”, “Centro-direita”, “Oportunidade”). Já o corpus online recorre sobretudo a: (i) eventos históricos ou traumáticos (“Impeachment”, “Massacre”, “República”); (ii) atores e antagonistas concretos (“Banqueiros”, “Milionários”, “Reacionários”, “Governadores”); (iii) operações econômicas da vida cotidiana (“Aluguel”, “Empréstimo”, “Financiamento”); e (iv) efeitos sociais materializados (“Homicídios”, “Pobreza”, “Desigualdade”). Não se trata de categorias exaustivas mas de uma codificação funcional suficientemente estável para mostrar que os cinco casos não são apenas cinco anedotas.

5.1 “Democracia” → “Ditadura”

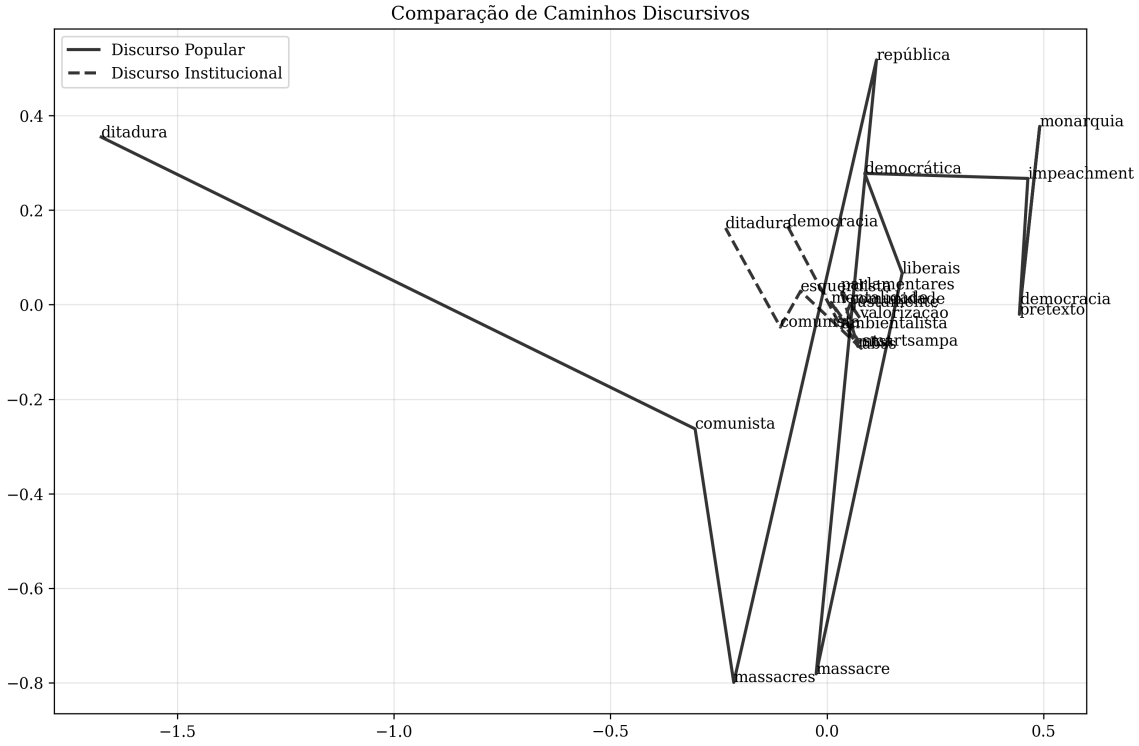

Figura 1: “Democracia” → “Ditadura”, gráfico comparativo entre discursos.

Debate	Democracia → Ambientalista → Comunidade → Justamente → Valorização → Parlamentares → MTST → Mentalidade → Smart Sampa → UBSs → Esquerdista → Comunista → Ditadura
Popular	Democracia → Monarquia → Pretexto → Impeachment → Democrática → Liberais → Massacre → República → Massacres → Comunista → Ditadura

O contraste local é nítido. No corpus do debate, a transição entre "democracia" e "ditadura" atravessa mediadores político-institucionais ("Parlamentares", "MTST") e, de modo ainda mais revelador, itens de gestão urbana ("Smart Sampa", "UBSs"). O antagonismo máximo da política moderna é conectado, portanto, por uma rota que o desloca para dentro do vocabulário administrativo. Mesmo quando surgem marcadores ideológicos mais explícitos, como "Esquerdista" e "Comunista", eles aparecem depois de uma longa zona de mediação burocrática.

No corpus online, o caminho é outro. A rota passa por "Monarquia", "Pretexto", "Impeachment", "Massacre", "República" e "Massacres", isto é, por uma sequência que historiciza o conflito e o ancora em eventos e formas políticas reconhecíveis. A ditadura não surge como abstração terminal, mas como memória de violência e ruptura. Alguns itens do caminho institucional inicial (como "Ambientalista", "Justamente" e "Valorização") são menos transparentes, e isso precisa ser dito. Ainda assim, os elos mais interpretáveis convergem para a mesma leitura: no debate, o conflito passa por instâncias de gestão; no corpus online, ele passa por memória histórica. Esse caso importa para o padrão geral porque mostra, já no primeiro par, que a diferença entre os corpora não é apenas temática, mas de regime de mediação.

5.2 “Dinheiro” → “Elite”

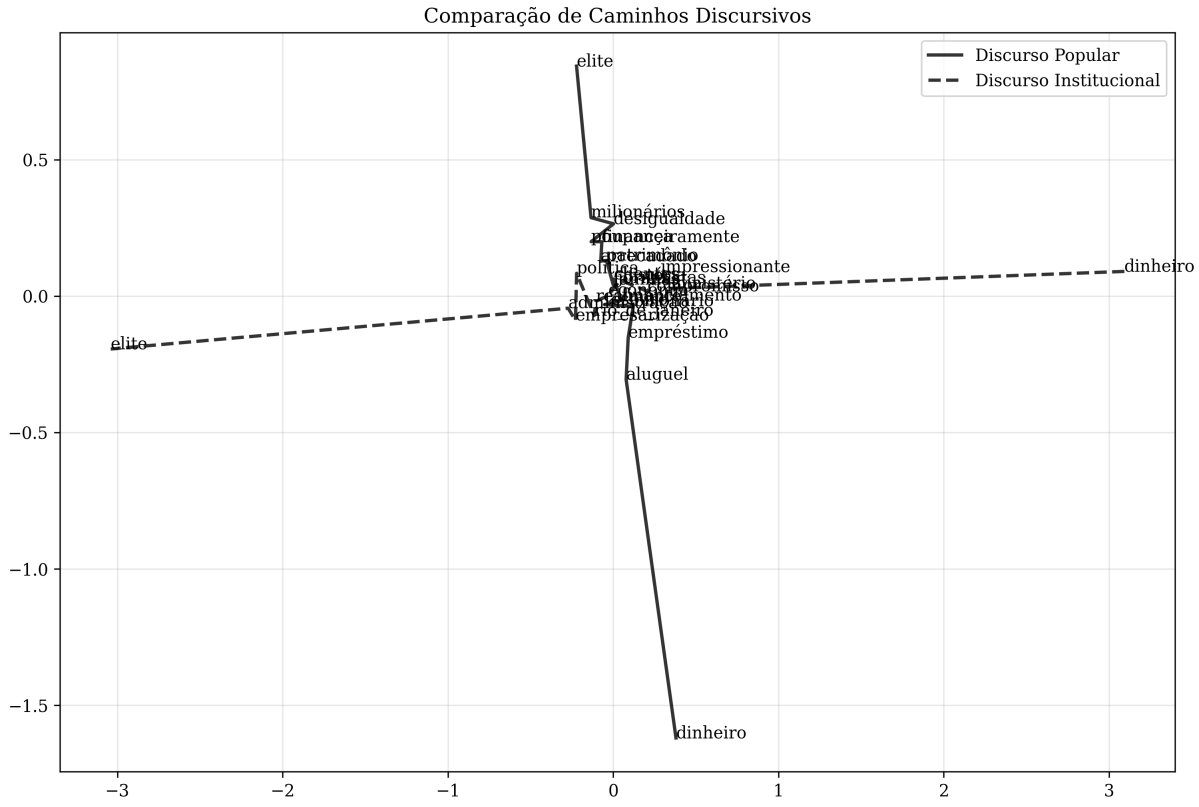

Figura 2: “Dinheiro” → “Elite”, gráfico comparativo entre discursos.

Debate	Dinheiro → Ministério → Impressionante → Compromisso → História → Jornalistas → Economia → Realmente → Rio de Janeiro → Política → Empresarização → Administração → Elite
Popular	Dinheiro → Aluguel → Empréstimo → Bilionário → Tesouro → Financiamento → Clientes → Famílias → Patrimônio → Arrecadado → Financeiramente → Poupança → Desigualdade → Milionários → Elite

Entre “dinheiro” e “elite”, o corpus do debate produz um percurso de legitimação mediada. “Ministério”, “Jornalistas”, “Política” e “Administração” enquadram a elite como resultado de instâncias formais, narrativas públicas e gestão. A presença de “Impressionante” logo após “Ministério” é especialmente sugestiva: o caminho não apenas administra a elite, mas a reveste de avaliação positiva. Há alguns itens menos estáveis semanticamente, como “Realmente”, mas o eixo do caminho permanece claro: a elite é alcançada por mediações institucionais e discursivas.

No corpus online, em contraste, a elite é construída a partir de operações econômicas vividas: “Aluguel”, “Empréstimo”, “Financiamento”, “Poupança”. A cadeia se desloca da circulação ordinária do dinheiro para a acumulação desigual de patrimônio, até chegar a “Desigualdade”, “Milionários” e “Elite”. O que localmente se observa é uma mudança de escala: o debate trata a elite como produto de ordenações formais; o corpus online a reconstrói a partir de práticas econômicas cotidianas e de seus efeitos sociais. Isso importa para o padrão mais amplo porque reforça a oposição entre uma mediação institucionalizante e uma mediação materialmente ancorada. Onde o debate abstrai, o corpus online concretiza.

5.3 “Governo” → “População”

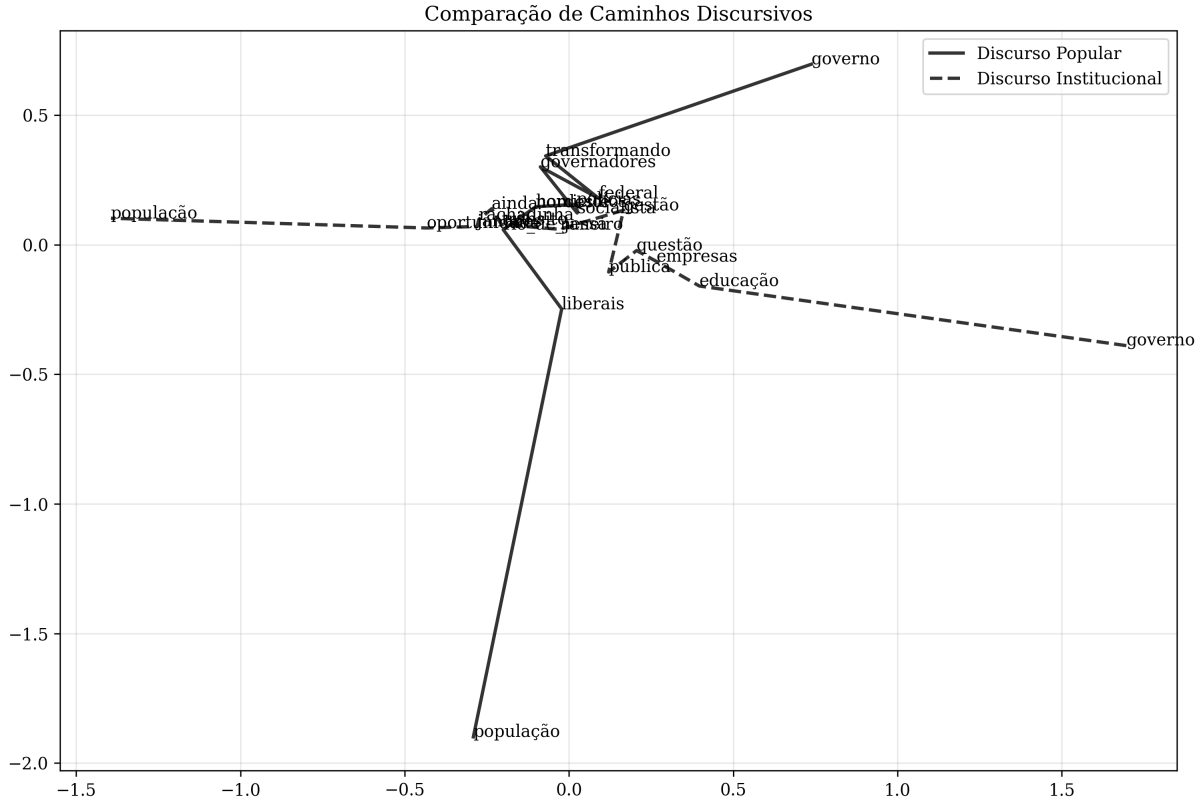

Figura 3: “Governo” → “População”, gráfico comparativo entre discursos.

Debate	Governo → Educação → Empresas → Questão → Pública → Gestão → Nessa → Trânsito → Ainda → Rachadinha → Janones → Oportunidade → População
Popular	Governo → Transformando → Federal → Governadores → Socialista → Polícias → Homicídios → Nordeste → Rio de Janeiro → Liberais → População

No corpus do debate, “governo” e “população” são ligados por um inventário de atribuições da gestão pública: “Educação”, “Questão Pública”, “Gestão”, “Trânsito”. Mesmo a controvérsia, quando aparece em “Rachadinha” e “Janones”, é absorvida pela cadeia administrativa e não reorganiza o trajeto. A população figura ao fim como destinatária de um conjunto de temas e problemas administráveis. Itens como “Nessa” e “Ainda” são pouco interpretáveis isoladamente, mas não desestabilizam a tendência dominante do caminho.

No corpus online, o mesmo par é conectado por um repertório mais áspero: “Federal”, “Governadores”, “Polícias”, “Homicídios”, “Nordeste”, “Rio de Janeiro”. O poder público aparece imediatamente distribuído em instâncias concretas, e a passagem por “Polícias” e “Homicídios” desloca a população do lugar de beneficiária da gestão para o de objeto de conflito e violência estatal. Também aqui há elos menos nítidos, como “Transformando”, mas o núcleo do caminho é semanticamente forte. O contraste local, portanto, é entre governo como administração de temas e governo como aparato cujos efeitos se materializam em violência, território e disputa ideológica. Nesse sentido, o encadeamento observado é compatível com achados recentes sobre a esfera política digital brasileira, em que “comportamentos verbais impolidos foram retribuídos com mais impolidez” (OLIVEIRA; MIRANDA, 2024, p. 494), fazendo com que repertórios de Estado, conflito e hostilidade se reforcem mutuamente na circulação online. No plano geral, o caso reforça que o corpus online não apenas “fala mais forte”; ele conecta os conceitos por consequências sociais mais tangíveis.

5.4 “Pobres” → “Elite”

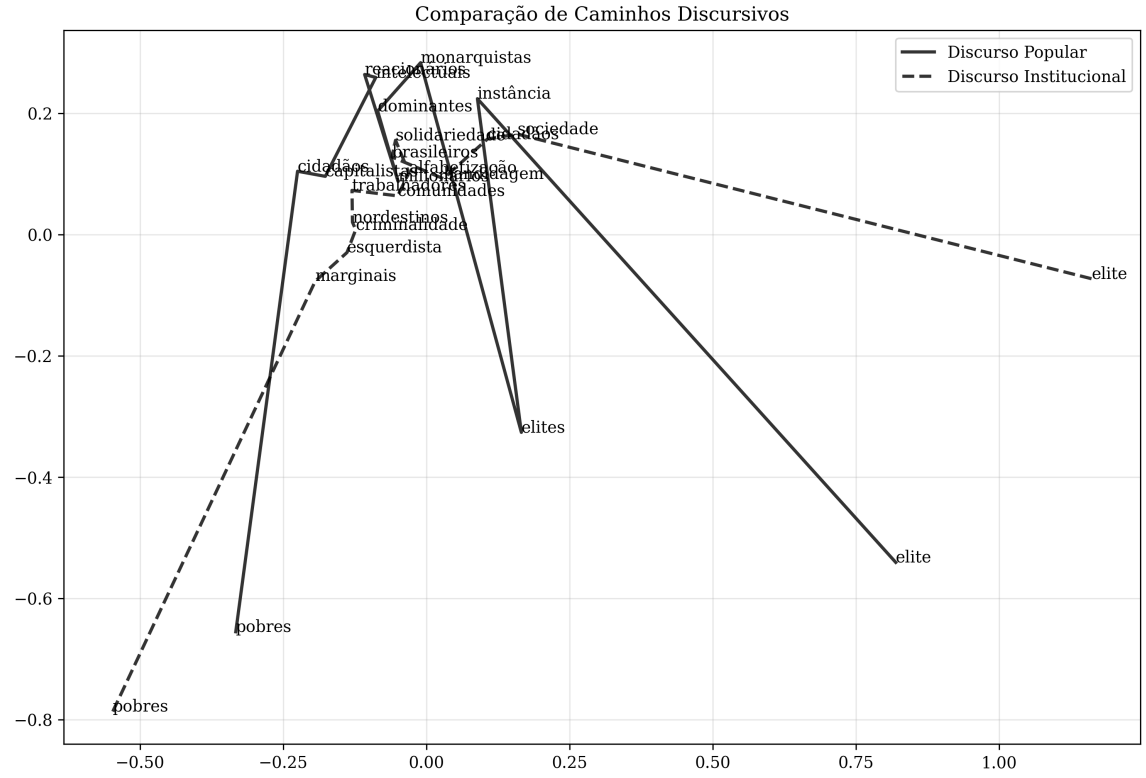

Figura 4: “Pobres” → “Elite”, gráfico comparativo entre discursos.

Debate	Pobres → Marginais → Esquerdista → Criminalidade → Nordestinos → Trabalhadores → Comunidades → Alfabetização → Solidariedade → Brasileiros → Bandidagem → Cidadãos → Sociedade → Elite
Popular	Pobres → Cidadãos → Capitalistas → Intelectuais → Reacionários → Milionários → Dominantes → Monarquistas → Elites → Instância → Elite

Este é o caso em que a assimetria entre os corpora se torna mais sensível. No debate, a transição de “pobres” para “elite” passa primeiro por “Marginais” e “Criminalidade”, e só então alcança “Nordestinos” e “Trabalhadores”. Em seguida, o caminho propõe uma integração social burocratizada por “Comunidades”, “Alfabetização”, “Solidariedade”, “Cidadãos” e “Sociedade”. O resultado é uma rota longa, sinuosa e internamente tensa: a mesma cadeia que integra também criminaliza. A passagem por “Nordestinos” entre “Criminalidade” e “Trabalhadores” merece destaque analítico, ainda que sem extrapolação causal além do corpus.

No corpus online, a lógica é outra. “Pobres” vai a “Cidadãos” e rapidamente a “Capitalistas”, “Reacionários”, “Milionários”, “Dominantes” e “Monarquistas”. Não há desvio pela

criminalização, nem promessa de integração mediada; há uma "representação antagonística de grupos sociais" (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 18) pela nomeação progressivamente mais explícita de antagonistas e posições de poder. "Instância" é um elo mais opaco, mas ele não impede a leitura do conjunto. O contraste local mostra um debate que dilui o conflito social em uma gramática de integração e administração, contra um corpus online que o reconstroi como antagonismo de classe. No padrão mais amplo, este talvez seja o caso que melhor evidencia a diferença entre burocratização semântica e nomeação direta do conflito.

5.5 "População" → "Desemprego"

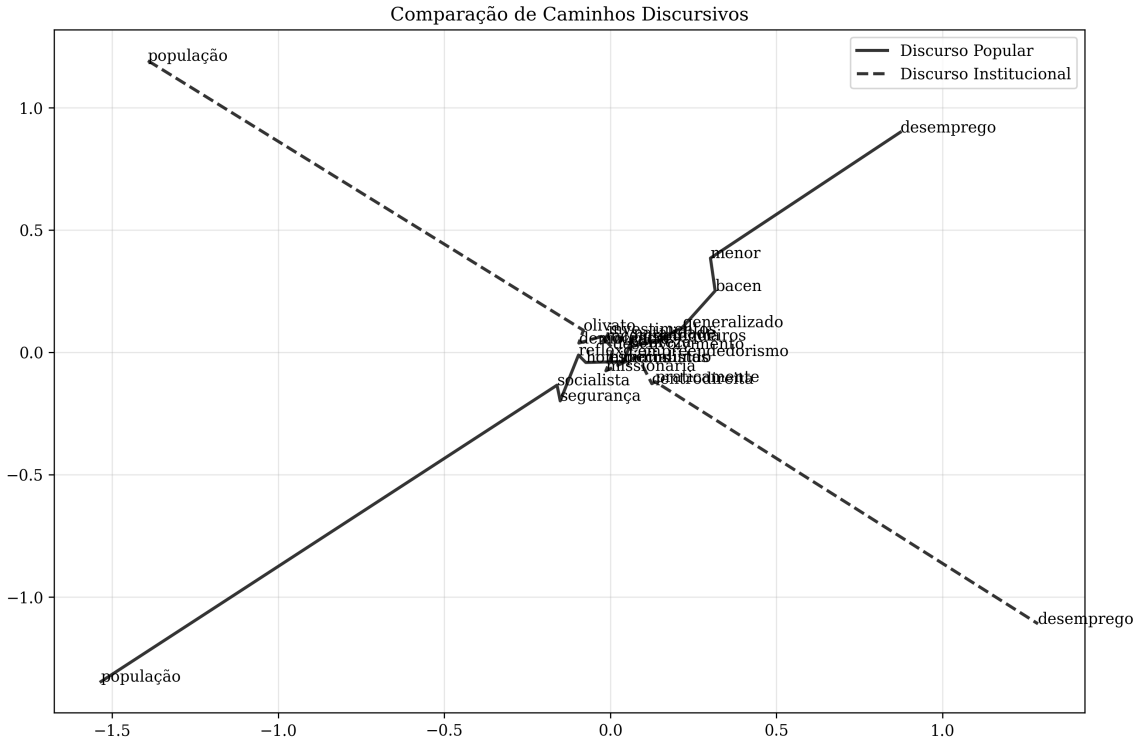

Figura 5: "População" → "Desemprego", gráfico comparativo entre discursos.

Debate	População → Olivato (Eduardo Olivatto) → Democracia → Investimentos → Violência → Incompetência → Desenvolvimento → Especialistas → Missionária → Empreendedorismo → Centro-direita → Praticamente → Desemprego
Popular	População → Socialista → Segurança → Reflexo → Homicídios → Diminuindo → Natalidade → Pobreza → Banqueiros → Generalizado → Bacen → Menor → Desemprego

O último par focaliza um ponto clássico do discurso político: a relação entre população e emprego. No corpus do debate, o trajeto passa por “Investimentos”, “Incompetência”, “Desenvolvimento”, “Especialistas”, “Empreendedorismo” e “Centro-direita”. A cadeia sugere uma causalidade gerencial: o desemprego é concebido como problema de desenvolvimento insuficiente, má condução e necessidade de solução especializada. Aqui aparecem também itens mais opacos ou idiossincráticos, como “Olivato”, “Missionária” e “Praticamente”; eles recomendam cautela, mas não anulam a inteligibilidade do núcleo semântico.

No corpus online, a rota é mais estrutural. “Segurança”, “Homicídios”, “Natalidade”, “Pobreza”, “Banqueiros” e “Bacen” articulam emprego, violência, reprodução social e sistema financeiro. O desemprego não aparece como falha administrativa pontual, mas como efeito distribuído de uma ordem econômica e social mais ampla. “Reflexo”, “Diminuindo” e “Menor” são elos menos densos, mas os intermediários substantivos sustentam a leitura. Esse caso importa para o padrão geral porque condensa a oposição central da seção: no debate, o problema social é filtrado por especialistas e soluções ideologicamente gerenciais; no corpus online, ele é reancorado em indicadores concretos e agentes reconhecíveis.

5.6 Recorrência, alcance e limites

Tomados em conjunto, os cinco pares focalizados mostram recorrência suficiente para sustentar uma generalização prudente. No corpus do debate, repetem-se intermediários de gestão, administração pública, institucionalidade e solução técnico-política. No corpus online de referência, repetem-se intermediários ligados a memória histórica, antagonistas de classe, economia cotidiana, violência e atores concretos; esse padrão ajuda a tratar os caminhos discursivos geométricos como “formas produtivas de nomeação e predicação” (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 12), mas também como traços de circuitos polêmicos em que há “uma intensificação tão severa do desacordo de opiniões que acaba fazendo com que o ‘outro’ se cale” (TEIXEIRA; DAMASCENO-MORAIS, 2023, p. 1576), e não apenas como listas lexicais isoladas. Essa recorrência não elimina a necessidade de cautela. Nem todo elo do caminho é igualmente transparente; alguns itens são ambíguos, raros ou pouco informativos, e por isso a leitura proposta deve privilegiar as zonas de maior densidade interpretativa, não cada palavra isoladamente — sobretudo porque ambientes digitais antagonísticos também comportam a “criação de representações nebulosas do outro” (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 13) e podem levar à “estagnação e não desenvolvimento de argumentos” (TEIXEIRA; DAMASCENO-MORAIS, 2023, p. 29). Ao mesmo tempo, a leitura de conjunto ganha força porque, em estudo recente sobre agressão política online, “as tentativas de vilipêndio moral dos membros do ju-

diciário se tornaram uma tendência comum nas redes sociais brasileiras" (OLIVEIRA et al., 2025, p. 14); por isso, o valor analítico destes caminhos não está em cada elo isolado, mas na recorrência de combinações semânticas que condensam antagonismo, avaliação e circulação conflitiva. Essa cautela interpretativa se torna ainda mais importante quando se considera que, em ecossistemas digitais polarizados, a desinformação raramente se apresenta de modo isolado, mas compõe "um verdadeiro cipoal que desafia a compreensão linear" (MENDES et al., 2025, p. 14). Nessa chave, os intermediários do corpus online podem ser lidos menos como etiquetas temáticas autossuficientes do que como traços de cadeias de circulação em que "as modalidades do fazer-crer e do fazer-sentir não se manifestem de forma separada, mas sobreposta" (MENDES et al., 2025, p. 18). Também não se deve tratar os cinco pares como prova exaustiva do espaço semântico dos corpora. Eles são casos focais, selecionados por sua relevância analítica e mantidos porque correspondem aos pares efetivamente fixados no procedimento.

Ainda assim, a seção já permite uma formulação empiricamente mais responsável do achado central do artigo: nos cinco pares focalizados, observa-se recorrência de intermediários técnico-administrativos e institucionalizantes no corpus do debate, em contraste com intermediários mais historicamente ancorados, socialmente concretos ou antagonísticos no corpus online de referência. As checagens complementares de estabilidade, sensibilidade ao tamanho do corpus e comparação com versões menos restritivas do procedimento são remetidas ao material suplementar. No corpo do artigo, o ponto decisivo é outro: não se trata apenas de dizer que os dois corpora "falam diferente", mas de mostrar que conectam os mesmos conceitos politicamente carregados por regimes distintos de mediação semântica.

6 Discussão

Os resultados permitem interpretar os caminhos discursivos como *rotas de mediação* no interior de dois campos simbólicos distintos. É nesse sentido que o conceito de tecnocratização do discurso, em Fairclough (2001), se torna produtivo para a leitura do corpus do debate. O artigo não pretende demonstrar, em sentido forte, toda a teoria social da tecnocratização, nem derivar dela uma explicação causal exaustiva do processo eleitoral. O que a análise mostra é algo mais circunscrito: nos cinco pares focalizados, o espaço semântico do debate tende a conectar conceitos politicamente carregados por meio de intermediários de administração, gestão, institucionalidade e legitimação burocrática. A utilidade de Fairclough, aqui, está em oferecer uma chave interpretativa para esse modo de mediação, e não um selo automático de verdade para qualquer ocorrência de léxico técnico.

Essa distinção é importante porque o achado do artigo não se reduz à presença de palavras especializadas no debate. Muitos discursos públicos usam vocabulário técnico sem, por isso, reorganizar de modo relevante suas transições semânticas. O ponto específico dos caminhos discursivos é outro: eles mostram que, na comparação aqui construída, termos ligados à gestão pública, à administração e à institucionalidade aparecem *entre* conceitos-âncora de alta densidade política. Em outras palavras, o debate não apenas fala tecnicamente; ele tende a *mediar tecnicamente* certas passagens semânticas. Quando isso ocorre de maneira recorrente, a leitura pela tecnocratização ganha plausibilidade: questões de conflito, antagonismo ou desigualdade passam a ser atravessadas por uma camada de linguagem que as reenquadra como problemas de gestão, ordenamento ou operação institucional.

No corpus online de referência, a lógica observada é diferente. Os caminhos ali obtidos tendem, com mais frequência, a acionar eventos, figuras políticas, antagonismos nomeáveis e circuitos de economia cotidiana. Em vez de uma mediação predominantemente técnico-administrativa, aparecem rotas mais densamente ancoradas em memória discursiva e em referenciações historicamente reconhecíveis; em certos pontos, a própria indeterminação da nomeação faz com que "a voz autoral não se comprometa diretamente com uma posição explícita" (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 14). Essa diferença não autoriza afirmar que o corpus online seja "mais verdadeiro", "mais autêntico" ou mais fiel à realidade social. Autoriza apenas dizer que, nesta comparação, ele conecta certos conceitos por mediações mais concretas, mais eventificadas ou mais diretamente antagonísticas. Com isso, o artigo sugere que os dois corpora não apenas dizem coisas diferentes: eles *organizam diferentemente* o trajeto semântico entre conceitos próximos do vocabulário político.

Mesmo sem reduzir o corpus online de referência a um corpus de impolidez, os resultados aqui obtidos entram em diálogo com pesquisas recentes sobre conflito verbal em plataformas digitais. Oliveira e Miranda mostram que, em ecologias políticas online, a circulação antagonística pode envolver "a acumulação de recursos simbólicos (verbais e não-verbais) nas mensagens provocativas e/ou ofensivas" (OLIVEIRA; MIRANDA, 2024, p. 493) e que esses recursos contribuem para disseminar ataques contra "indivíduos ou instituições, retratados como inimigos" (OLIVEIRA; MIRANDA, 2024, p. 495). Essa formulação ajuda a qualificar a diferença observada neste artigo: no corpus contrastivo, a mediação semântica passa com mais frequência por atores, episódios e alvos nomeáveis, isto é, por cadeias em que o político reaparece menos como problema de gestão e mais como campo de conflito, imputação e antagonização.

Em chave complementar, Oliveira et al. mostram que, em ambientes de agressão política digital, "ao remover os limites de conduta verbal polida, licenciar ou até defender a

impolidez, o despudor verbal se torna uma prática importante" (OLIVEIRA et al., 2025, p. 7). Isso ajuda a interpretar o corpus online não só como repertório de menções mais concretas, mas como espaço em que o léxico pode operar como "estratégia deliberada de manipulação e manutenção do discurso" (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 7) e em que a nomeação antagonística pode funcionar como técnica de desqualificação. Mais que um desvio periférico do debate público, tal dinâmica importa porque "esse ciclo de agressão afeta o tecido social como um todo, e não apenas as mulheres, ou as pessoas LGBT+" (OLIVEIRA et al., 2025, p. 14). Nesse sentido, a diferença entre os corpora não é apenas de estilo ou de topicalidade: ela envolve, no corpus online, "práticas antagonísticas voltadas à criação discursiva de um eu em oposição a um outro" (PIVETTA; GONÇALVES-SEGUNDO, 2023, p. 6).

Essa leitura depende diretamente da decisão de tratar o debate eleitoral como gênero e, portanto, como unidade analítica própria. Ao reunir em um único espaço vetorial as falas de candidatos politicamente adversários, o artigo não pretende negar diferenças entre estilos, ideologias ou estratégias individuais. Essas diferenças existem e podem, em outros desenhos de pesquisa, tornar-se o próprio objeto central da análise. Aqui, porém, elas são deliberadamente *bracketed* para que se examine uma pergunta de outro nível: que tipo de campo simbólico emerge quando se toma o debate eleitoral como prática discursiva relativamente estável, com regras de circulação, expectativas de performance e formas reconhecíveis de mediação? Nesse enquadramento, os candidatos são adversários no processo político, mas colaboradores na constituição do gênero debate. O *pooling* das falas, portanto, não apaga vozes por descuido; ele as suspende analiticamente para tornar visível uma propriedade de nível genérico.

Algo semelhante vale para o segundo corpus. O artigo não depende da hipótese de que os comentários coletados representem a população brasileira em qualquer sentido sociológico forte. Trata-se, antes, de um *corpus de comentários online de referência*, usado como interlocutor contrastivo para o campo simbólico do debate. Essa formulação é decisiva, porque desloca a comparação de um registro representacional para um registro relacional. Não se pergunta aqui "como o povo brasileiro fala", mas sim se um corpus institucional e um corpus online heterogêneo conectam os mesmos conceitos por mediações semelhantes ou distintas. O valor analítico do segundo corpus, portanto, é contrastivo: ele permite observar como certas transições semânticas se reorganizam quando mudam o gênero, a situação de enunciação, a escala temática e as condições de produção do discurso. Essa cautela se reforça quando lembramos que, nas mídias sociais, as mensagens "circulam para uma audiência abstrata, composta quase em sua totalidade por desconhecidos" (OLIVEIRA et al., 2025, p. 15), o que torna menos nítidos os contornos pragmáticos e sociológicos do público efetivo desse material.

Nessa mesma direção, Mendes et al. lembram que, em ecossistemas digitais marcados por circulação desinformacional, "dificilmente a desinformação (entendida como conteúdo enganoso/mentiroso) se apresenta de forma pura" (MENDES et al., 2025, p. 14). A observação é útil aqui porque impede que o segundo corpus seja lido como bloco homogêneo ou como repertório semanticamente unívoco: em materiais desse tipo, "uma mesma postagem pode ter elementos de uma ou mais tipologias" (MENDES et al., 2025, p. 15) e, mais ainda, tais tipologias são "categorias abstratas e podem coexistir em uma mesma textualidade" (MENDES et al., 2025, p. 18). Para o presente artigo, isso reforça uma cautela metodológica decisiva: os comentários online importam menos como espelho transparente de posições sociais estáveis do que como superfície híbrida em que crença, afeto, antagonização e circulação se sobrepõem.

Nessa mesma direção, o estudo de Tucci e Gouveia sobre grupos pró-Bolsonaro no Telegram ajuda a refinar a leitura do corpus contrastivo quando a circulação discursiva depende de simplificação estratégica e de forte centralização de canais. Os autores observam que, mesmo em mensagens problemáticas, "a tônica dominante na formulação do discurso privilegiou a não utilização de estratégias comunicativas excessivamente provocativas" (TUCCI; GOUVEIA, 2025, p. 16), o que impede identificar a circulação antagonística apenas por marcas máximas de agressividade lexical. Ao mesmo tempo, mostram que o canal oficial de Bolsonaro "utilizou sua plataforma para disseminar conteúdos tendenciosos e alimentar um clima de desconfiança" (TUCCI; GOUVEIA, 2025, p. 18) e que o então presidente "capitalizou sua posição para ampliar seu alcance comunicativo na plataforma" (TUCCI; GOUVEIA, 2025, p. 21). Para o presente artigo, isso é relevante porque reforça que caminhos discursivos mais eventificados ou mais ancorados em atores nomeáveis não precisam assumir sempre a forma de insulto explícito: eles podem operar também por campanhas de simplificação temática, coordenação de circulação e ativação de desconfiança.

Nessa chave, a contribuição do presente artigo não está em substituir análises pragmáticas finas do conflito verbal, mas em acrescentar uma camada mesoestrutural à interpretação. A métrica proposta não identifica, por si só, atos de impolidez, mas ajuda a localizar cadeias em que a organização semântica do corpus favorece a reiteração de antagonistas, acusações e alvos recorrentes. Isso se aproxima do ponto em que, para Oliveira e Miranda, "o metadiscurso revela que/como os comportamentos verbais, classificados como impolidos, são retaliados" (OLIVEIRA; MIRANDA, 2024, p. 495). Aqui, a vantagem é mostrar que essa lógica pode deixar vestígios não apenas em enunciados isolados, mas também na topologia relacional dos caminhos entre conceitos politicamente densos.

É nesse ponto que a noção de *autoproteção semântica* pode ser introduzida com

cautela. O que aqui chamamos, de forma provisória, de autoproteção semântica não designa uma intenção consciente dos falantes nem um mecanismo totalizante do discurso institucional. Designa, mais modestamente, um efeito observável na organização dos caminhos: associações politicamente sensíveis tendem a aparecer amortecidas por mediadores institucionalmente legítimos, o que dificulta sua formulação em termos diretos. A expressão é útil porque nomeia um possível efeito de distanciamento semântico, mas deve permanecer provisória. Em corpora maiores, em outros gêneros e com outros pares-âncora, esse efeito poderá aparecer de modo diferente ou mesmo não aparecer.

Alcance e limites da comparação

O alcance probatório do artigo deve ser formulado com a mesma precisão. Os cinco pares focalizados não esgotam o espaço semântico dos corpora nem autorizam universalizações fortes sobre todo discurso institucional ou sobre todo comentário online. As checagens de robustez resumidas no artigo e detalhadas no material suplementar ampliam a confiança na recorrência do contraste, mas não encerram a questão. O que esta pesquisa permite afirmar, com segurança razoável, é que a métrica dos caminhos discursivos torna visíveis diferenças sistemáticas de mediação semântica entre dois campos simbólicos construídos em condições de produção distintas. O que ela ainda não permite afirmar é que essas diferenças esgotem a heterogeneidade interna dos corpora, que valham igualmente para qualquer par de conceitos ou que descrevam, sem mediações adicionais, totalidades sociais mais amplas. Ainda assim, esse resultado constitui contribuição relevante para a Linguística textual e discursiva: ele oferece uma forma nova de descrever como o texto organiza percursos entre conceitos e mostra que tais percursos podem ser interpretados teoricamente sem reduzir a análise discursiva à técnica nem prescindir de transparência metodológica.

7 Conclusão

Este artigo apresentou a métrica dos caminhos discursivos geométricos como instrumento de análise contrastiva entre dois campos simbólicos treinados separadamente. Sua contribuição metodológica está no critério de persistência: ao exigir que os intermediários se mantenham ao longo de múltiplas projeções do espaço vetorial, a análise privilegia elos mais robustos e reduz o peso de aproximações ocasionais produzidas por uma única perspectiva dimensional.

No plano empírico, a comparação aqui construída mostrou, nos cinco pares analisa-

dos, uma recorrência significativa: o corpus do debate tende a conectar conceitos politicamente densos por mediações institucionalmente legítimas ou técnico-administrativas, ao passo que o corpus de comentários online de referência tende a ativar intermediários mais diretamente ancorados em eventos, antagonistas nomeados e operações da vida econômica cotidiana. Essa diferença não autoriza generalizações sobre todo discurso institucional nem sobre a sociedade brasileira; vale, antes, como resultado de um estudo de caso contrastivo sustentado por um conjunto focal de pares-âncora e por checagens de robustez resumidas no artigo e detalhadas no suplemento.

Em termos disciplinares, o trabalho sugere que uma métrica vetorial pode ser incorporada a uma pergunta de Linguística textual sem que a interpretação discursiva seja substituída pela técnica. O procedimento não lê o texto no lugar do analista; ele torna mais visíveis certas rotas de mediação semântica que depois precisam ser interpretadas teoricamente.

Como agenda futura, interessa ampliar o teste para corpora maiores, outros gêneros institucionais e desenhos comparativos mais balanceados, além de refinar limiares de persistência e procedimentos de robustez. O achado, assim, deve ser lido como uma proposta analítica já sustentada por evidência contrastiva, mas ainda aberta a expansão e revisão.

Conflito de interesses

O autor declara não haver conflitos de interesse financeiros, pessoais, acadêmicos, políticos ou comerciais que possam ter influenciado o trabalho reportado neste artigo.

Nota ética e normativa

Nos termos da Resolução CNS n.º 510/2016, art. 1º, parágrafo único, II, e do Ofício Circular n.º 17/2022/CONEP/SECNS/MS, o estudo enquadra-se na hipótese de dispensa de registro e avaliação pelo Sistema CEP/Conep, por utilizar exclusivamente informações de acesso público, sem interação com participantes, sem coleta de dados diretamente obtidos de pessoas e sem tratamento de informações identificáveis. O corpus online foi constituído apenas por comentários publicamente acessíveis, sem coleta ou armazenamento de nomes de usuário, identificadores de perfil ou outros dados pessoais. No caso dos debates, o repositório disponibiliza apenas as transcrições textuais e os artefatos analíticos necessários à reprodutibilidade; os arquivos brutos de áudio e vídeo das transmissões não são redistribuídos, por não serem necessários à reprodução da análise textual e por estarem sujeitos a restrições autorais e de licenciamento das emissoras e das plataformas de hospedagem. Todos os materiais utilizados, portanto, são materi-

ais textuais de acesso público. Sem prejuízo dessa dispensa no âmbito do Sistema CEP/Conep, permanecem resguardados eventuais fluxos institucionais próprios, quando aplicáveis.

Declaração de disponibilidade de dados

Os dados, materiais e rotinas analíticas que sustentam os resultados deste artigo estão disponibilizados em repositório aberto. O conjunto inclui os corpora textuais utilizados na análise, os arquivos necessários à sua reconstrução metodológica e os materiais suplementares associados ao estudo. Endereço do repositório: https://github.com/alexandre-barroso/GeometriaDiscursiva. O identificador persistente (DOI) deste repositório é https://doi.org/10.5281/zenodo.19392318.

Referências bibliográficas

ARAUJO, J. **O algoritmo é um texto**. **Texto Livre**, Belo Horizonte-MG, v. 18, p. e58505, 2025. DOI: 10.1590/1983-3652.2025.58505.

BACHINI, N.; BATISTA PILAU, L. Populismo penal ou digital? discursos da direita brasileira no Facebook. **Caderno CRH**, v. 38, p. e025078, 2025. DOI: 10.9771/ccrh.v38i0.67186.

BOURDIEU, P. **O poder simbólico**. Tradução de Fernando Tomaz. Lisboa: Edições 70, 2016.

BÜHLER, K. **Sprachtheorie: Die Darstellungsfunktion der Sprache**. Stuttgart: Gustav Fischer, 1965.

CHEN, J.; MIZUNO, T.; DOI, S. Analyzing political party positions through multi-language twitter text embeddings. **Frontiers in Big Data**, v. 7, p. 1330392, 2024. DOI: 10.3389/fdata.2024.1330392.

DIESSEL, H. Bühler's two-field theory of pointing and naming and the deictic origins of grammatical morphemes. In: BREBAN, T.; BREMS, L.; DAVIDSE, K.; MORTELMANS, T. (Eds.). **New perspectives on grammaticalization: Theoretical understanding and empirical description**. Amsterdam: John Benjamins, 2012. p. 35–48.

FAIRCLOUGH, N. **Discurso e mudança social**. Tradução de Izabel Magalhães. Brasília: Edi-

tora UnB, 2001.

FREDÉN, A.; JOHANSSON, M.; SAYNOVA, D. Word embeddings on ideology and issues from Swedish parliamentarians' motions: a comparative approach. **Journal of Elections, Public Opinion and Parties**, 2024. DOI: 10.1080/17457289.2024.2433979.

FREITAS, A. L.; ROMERO, R. L.; PANTALEÃO, F. N.; BOGGIO, P. S. Bases sociocognitivas do discurso de ódio online no Brasil: uma revisão narrativa interdisciplinar. **Texto Livre**, Belo Horizonte-MG, v. 16, p. e46002, 2023. DOI: 10.1590/1983-3652.2023.46002.

GRAVE, E.; BOJANOWSKI, P.; GUPTA, P.; JOULIN, A.; MIKOLOV, T. Learning word vectors for 157 languages. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki: European Language Resources Association, 2018. p. 3483–3487.

HANKS, W. F. **Língua como prática social: das relações entre língua, cultura e sociedade a partir de Bourdieu e Bakhtin**. Tradução de Anna Christina Bentes, Renato C. Rezende e Marco Antônio Rosa Machado. São Paulo: Cortez, 2008.

ITUASSU, A.; PECORARO, C.; CAPONE, L.; LEO, L.; MANNHEIMER, V. Mídias digitais, eleições e democracia no Brasil: uma abordagem qualitativa para o estudo de percepções de profissionais de campanha. **Dados**, Rio de Janeiro, v. 66, n. 2, p. e20210063, 2023. DOI: 10.1590/dados.2023.66.2.294.

KOCH, I. G. V. **Introdução à Linguística Textual: trajetórias e grandes temas**. 3. ed. São Paulo: Martins Fontes, 2013.

KRYKONIUK, K.; HOPKIN-KING, C.; ROBERTS, S. G. Developing a discourse space for analysing online discourse. **Discourse, Context & Media**, v. 67, p. 100929, 2025. DOI: 10.1016/j.dcm.2025.

MENDES, C. M.; ALMEIDA, C. G. de; MATTOS, M. Â. Ecossistema desinformacional em canais de extrema-direita no Telegram sobre as enchentes no RS em 2024: abordagem semiótico-interacional. **Texto Livre**, Belo Horizonte-MG, v. 18, p. e58644, 2025. DOI: 10.1590/1983-3652.2025.58644.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

OLIVEIRA, A. L. A. M.; MIRANDA, M. V. O metadiscurso de impolidez como recurso analítico: evidências do domínio político no Twitter/X. **Revista de Estudos da Linguagem**, Belo Horizonte, v. 32, n. 2, p. 479–498, 2024. DOI: 10.17851/2237-2083.32.2.479-498.

OLIVEIRA, A. L. A. M.; MIRANDA, M. V.; DRINÓCZI, T.; CUNHA, G. X. As novas fronteiras do dizível: a linguagem impolida contra figuras femininas das esferas política e judiciária. **Texto Livre**, Belo Horizonte-MG, v. 18, p. e58111, 2025. DOI: 10.1590/1983-3652.2025.58111.

OLIVEIRA, A. N. C. Democracia, populismo e discurso do voto impresso: análise de conteúdo no Facebook por mineração de texto e redes semânticas. **Dados**, Rio de Janeiro, v. 67, n. 4, p. e20220148, 2024. DOI: 10.1590/dados.2024.67.4.330.

PEREIRA E SILVA, W. Argumentação em discursos de ódio no Facebook: uma categorização contributiva à Linguística Forense e à Linguística Computacional. **Revista de Estudos da Linguagem**, Belo Horizonte, v. 29, n. 4, p. 2367–2395, 2024. DOI: 10.17851/2237-2083.29.4.2367-2395.

PIVETTA, L. M.; GONÇALVES-SEGUNDO, P. R. Anões, chads e nazipardos: as estratégias de nomeação e predicação no discurso da direita-alternativa no Brasil. **Texto Livre**, Belo Horizonte-MG, v. 17, p. e46380, 2024. DOI: 10.1590/1983-3652.2024.46380.

RUYTERS, W.; VERMEER, S.; KRUIKEMEIER, S.; VLIEGENTHART, R. Embedding, embedding on the wall; who is the most prominent politician of them all? Exploring automated methods to study multimodal political news coverage. **Communication Methods and Measures**, v. 20, n. 1, p. 26–52, 2026. DOI: 10.1080/19312458.2025.2558736.

TEIXEIRA, L. B. de S.; DAMASCENO-MORAIS, R. Espraiamento da estase argumentativa em interações polêmicas do Twitter. **Revista de Estudos da Linguagem**, v. 31, n. 3, p. 1557–1589, 2024. DOI: 10.17851/2237-2083.31.3.1557-1589.

TUCCI, G.; GOUVEIA, F. C. O discurso do bolsonarismo nas eleições 2022: uma investigação da desinformação viral em grupos de Telegram. **Em Questão**, Porto Alegre, v. 31, p. e141754, 2025. DOI: 10.1590/1808-5245.31.141754.

Apêndice

A Material suplementar

A.1 Nota de leitura e correspondência com o artigo

Este material suplementar concentra o detalhamento metodológico e documental deslocado do corpo do artigo por economia editorial. Sua função não é introduzir uma tese paralela, mas tornar auditável o procedimento que sustenta a análise principal. Em coerência com o contrato do manuscrito, mantemos em primeiro plano a pergunta de Linguística textual e discursiva, tratando a modelagem vetorial como instrumento analítico e não como fim em si mesma.

O suplemento cobre cinco frentes: (i) a especificação analítica integral dos caminhos discursivos geométricos; (ii) a construção e a preparação dos corpora; (iii) as rotinas de controle de qualidade aplicadas aos modelos vetoriais; (iv) as verificações complementares de robustez e os critérios de leitura adotados; e (v) a apresentação expandida dos cinco pares focais, com atenção explícita a itens opacos ou idiossincráticos. Em termos de correspondência com o artigo, esta seção complementa sobretudo a metodologia, o corpus, os resultados e a discussão.

A.2 Especificação analítica integral

Recurso inicial, treinamento e comparação entre espaços

A análise parte de vetores *fastText* pré-treinados para o português do *Common Crawl*, utilizados como ponto de partida para dois treinamentos independentes de Word2Vec. A comparação, portanto, não foi construída em um único espaço comum, mas em dois espaços vetoriais ajustados separadamente: um para o corpus dos debates e outro para o corpus de comentários online de referência. Essa decisão decorre do objetivo do artigo: comparar a forma das mediações semânticas produzidas por cada corpus, e não medir distâncias absolutas em um espaço unificado.

Em ambos os modelos, a dimensionalidade foi fixada em 300. Também permaneceram constantes, entre os corpora, a amostragem negativa (*negative*=15), a taxa inicial de aprendizado ($\alpha = 0.025$), a taxa mínima ($\alpha_{\min} = 0.0001$), o número de épocas (10) e o procedimento de ajuste sobre o vocabulário efetivamente presente em cada corpus. O treinamento foi inicializado pela sobreposição entre o vocabulário do corpus e os vetores pré-treinados disponíveis, de modo que itens já contemplados pelo recurso inicial partissem de representações

lexicalmente informadas antes do ajuste ao material específico do estudo.

A especificação efetivamente utilizada difere, porém, entre os dois corpora. No corpus dos debates, o modelo foi treinado com arquitetura CBOW, janela de contexto 7 e frequência mínima 2. No corpus de comentários online, utilizou-se arquitetura Skip-gram, janela 10 e frequência mínima 8. Essa assimetria precisa ser assumida explicitamente: ela faz parte do procedimento preservado. No desenho deste artigo, essa diferença não inviabiliza a comparação porque o objeto comparado não é um valor absoluto de proximidade entre corpora, mas o tipo de rota intermediária que cada espaço produz para os mesmos pares de âncoras.

Da preparação textual ao pré-processamento para modelagem

A preparação dos corpora ocorreu em duas camadas. A primeira foi uma limpeza textual prévia, voltada à remoção de artefatos evidentes de coleta e transcrição. A segunda foi um pré-processamento lexical aplicado já na etapa de treinamento vetorial. Essa distinção é importante: nem toda decisão tomada para tornar o texto legível coincide com a filtragem efetivamente usada para construir o espaço distribuicional.

Na etapa de modelagem, o texto foi tokenizado em português e convertido para caixa baixa. Em seguida, aplicou-se uma rotina de filtragem lexical com quatro decisões centrais. Primeiro, removeu-se o grosso das *stopwords* do português. Segundo, preservaram-se deliberadamente algumas palavras funcionais com potencial valor discursivo, como *não*, *sim*, *muito*, *pouco*, *mais* e *menos*. Terceiro, manteve-se apenas material compatível com um padrão alfabético do português, a fim de reduzir ruído gráfico residual. Quarto, conservaram-se alguns itens monolíteros de alta recorrência textual, como *é*, *à*, *e*, *o* e *a*. Não houve lematização, etiquetagem morfossintática nem reconhecimento automático de entidades nomeadas. O procedimento é lexical-distribucional e opera sobre coocorrência.

As linhas vazias foram descartadas, e apenas unidades com mais de três tokens após o pré-processamento foram efetivamente aproveitadas no treinamento. Essa decisão define a granularidade do material sobre o qual o modelo aprende regularidades de contexto: não se trata de uma análise por período sintático plenamente normalizado, mas de sequências textuais curtas o bastante para preservar localidade de coocorrência e longas o bastante para não degenerarem em ruído mínimo.

Extração dos caminhos discursivos geométricos

Dado um par de conceitos-âncora, a rotina de extração do caminho discursivo começa por recuperar os vetores correspondentes aos dois termos no espaço já treinado. Em seguida, calcula-se a média global do espaço vetorial e centralizam-se todos os vetores em torno dessa média. Sobre esse espaço centralizado, calcula-se uma decomposição por componentes principais com semente fixa (*random_state*=42), gerando uma base ortogonal completa para projeções sucessivas do campo semântico.

O cálculo da persistência percorre projeções dimensionais decrescentes do espaço, de 300 dimensões até 2 dimensões. Em cada projeção, calcula-se o ponto médio entre os dois âncoras projetados e selecionam-se os dez itens mais próximos desse ponto. A proximidade ao ponto médio, contudo, não basta para que um item seja aceito como intermediário. Para evitar que palavras lateralmente próximas ao centro, mas exteriores ao segmento semântico relevante, entrem no caminho, cada candidato ainda precisa satisfazer uma condição geométrica: sua projeção deve cair efetivamente entre os dois âncoras. Em termos operacionais, isso corresponde ao critério $0 < \textit{proj\_len} < \textit{path\_len}$.

Cada vez que um vocábulo atende a esse critério em uma dada projeção, ele acumula um ponto de persistência. Ao fim do percurso dimensional, mantêm-se apenas os itens cuja persistência ultrapassa 30% do número total de projeções consideradas. Esse limiar é a contribuição metodológica central do artigo. Seu papel não é identificar qualquer vizinho local do ponto médio, mas selecionar palavras que se mantêm como intermediárias sob mudanças sucessivas de perspectiva dimensional. Em outras palavras, o caminho final não corresponde a uma cadeia gananciosa de vizinhança, mas a uma sequência ordenada de mediações robustas.

A ordenação final dos intermediários preservados é calculada por sua posição média projetada em subconjuntos mais altos do espaço. Na rotina preservada, essa média é obtida a partir das projeções em 300, 250, 200, 150 e 100 dimensões. O resultado é uma sequência orientada entre os dois polos semânticos, lida no artigo como *rota de mediação* entre conceitos politicamente relevantes.

Casos-limite, empates e critérios de exclusão

O primeiro caso-limite potencial seria a ausência de um dos âncoras no vocabulário do modelo treinado. Esse problema não ocorreu entre os cinco pares focais do artigo. O segundo caso-limite envolve empates ou proximidades muito estreitas entre candidatos intermediários. Como a rotina seleciona os dez itens mais próximos do ponto médio em cada projeção e depois

submete esses itens ao filtro geométrico e à regra de persistência, pequenos empates locais têm importância menor do que a recorrência global do candidato ao longo das projeções. O suplemento não trata, portanto, a estabilidade do procedimento como identidade literal de cada posição em cada execução, mas como manutenção do núcleo funcional do caminho.

Um terceiro tipo de caso-limite é a presença de itens semanticamente opacos, idiossincráticos ou de leitura fraca. Eles não são eliminados a posteriori para "embelezar" o resultado. Pelo contrário: permanecem registrados e são discutidos nas tabelas suplementares. O critério interpretativo adotado no artigo privilegia núcleos de maior densidade semântica, sem fingir que todos os intermediários têm a mesma transparência discursiva. Essa distinção entre itens fortemente interpretáveis e itens mais opacos é parte da honestidade analítica do procedimento.

Quadro expandido de especificação

Item	Especificação
Recurso inicial	Vetores *fastText* pré-treinados para o português do *Common Crawl*.
Dimensionalidade	300 em ambos os modelos.
Treinamento	Word2Vec com ajuste separado por corpus.
Arquitetura	Debates: CBOW; comentários: Skip-gram.
Janela	Debates: 7; comentários: 10.
Frequência mínima	Debates: 2; comentários: 8.
Épocas	10.
Amostragem negativa	15.
Taxa de aprendizado	Inicial 0.025; mínima 0.0001; decaimento linear ao longo das épocas.
Tokenização	Tokenização em português; caixa baixa; filtragem lexical.
Stopwords	Removidas com preservação deliberada de *não*, *sim*, *muito*, *pouco*, *mais* e *menos*.
Filtros adicionais	Padrão alfabético do português; preservação de *é*, *à*, *e*, *o*, *a*; descarte de linhas vazias e de unidades com menos de quatro tokens.
Base de projeção	PCA calculada sobre o espaço centralizado.
Semente fixa	*random_state*=42 para a PCA.
Projeções para persistência	Dimensões decrescentes de 300 até 2.
Candidatos por projeção	Dez itens mais próximos do ponto médio entre os âncoras.
Regra geométrica de inclusão	Apenas candidatos cuja projeção recaia entre os dois âncoras.
Regra de persistência	Manutenção de itens com ocorrência superior a 30% das projeções.
Ordenação final	Posição média projetada em 300, 250, 200, 150 e 100 dimensões.
Pares-âncora	democracia–ditadura; dinheiro–elite; governo–população; pobres–elite; população–desemprego.
Escopo interpretativo	Comparação entre rotas de mediação em dois espaços treinados separadamente.

Tabela 4: Quadro expandido de especificação analítica

A.3 Construção e preparação dos corpora

Corpus dos debates

O corpus institucional do artigo reúne debates televisivos e digitais da eleição para a prefeitura de São Paulo em 2024. Entraram no conjunto os debates da Band (primeiro turno), da Record (primeiro e segundo turnos), da TV Cultura (primeiro turno), do Terra (primeiro turno) e do UOL (primeiro turno), perfazendo 13h04m13s de material. O objetivo não foi construir um banco de falas individuais por candidato, mas um texto único representativo do debate eleitoral enquanto gênero. Por essa razão, as transcrições finais foram agregadas sem marcação de falante.

A construção desse corpus ocorreu por transcrição automática em segmentos sucessivos de 60 segundos, processados em português com modelo Whisper *medium*. A segmentação temporal em blocos curtos teve função prática: reduzir o custo de processamento, manter a continuidade em arquivos longos e tornar auditável a passagem entre o áudio original e o texto consolidado. Em seguida, os segmentos foram concatenados na ordem original do debate, preservando a linearidade textual do evento. Essa rotina mostrou-se especialmente adequada a um material com múltiplos falantes, sobreposições, interrupções e variações importantes de volume e clareza sonora.

Depois da transcrição, o corpus dos debates passou por uma limpeza textual em duas etapas. A primeira removeu marcas residuais do próprio processo de transcrição: rótulos de segmento, marcas temporais, URLs, numerais isolados e caracteres não linguísticos claramente estranhos ao corpus. Também houve uniformização para caixa baixa e normalização de espaços em branco. A segunda etapa realizou uma intervenção lexical mais pontual, voltada à remoção de resíduos recorrentes de plataforma ou transmissão e à regularização de algumas expressões multiword politicamente relevantes, como topônimos frequentes. Essa combinação entre limpeza ampla e correção seletiva foi importante para evitar que ruído mecânico se convertesse em coocorrência semântica artificial.

Além dessas intervenções, realizou-se correção manual de erros evidentes, sobretudo em nomes próprios, topônimos e itens cuja grafia alterava de maneira relevante a interpretabilidade política do corpus. O caso exemplar mencionado no projeto é a correção de formas como “De Atena” para “Datena”. O objetivo dessa etapa não foi estilizar o texto, mas impedir que falhas automáticas grosseiras contaminassem a distribuição lexical do espaço vetorial.

Corpus de comentários online de referência

O segundo corpus reúne 148.792 comentários públicos coletados no Reddit e agregados em um único arquivo textual. Seguindo a lógica adotada no corpo do artigo, este material é definido como *corpus de comentários online de referência*. A expressão "discurso popular", quando aparece, deve ser entendida apenas como abreviação operacional para esse corpus, e não como pretensão de representatividade demográfica do país.

A coleta buscou heterogeneidade temática e discursiva. Foram contempladas comunidades voltadas a política, economia, cotidiano, frustração pessoal e assuntos locais: *Brasil*, *BrasildoB*, *Brasilivre*, *Conversas*, *Desabafos*, *Investimentos* e *SaoPaulo*. A rotina preservada operava, por comunidade, sobre conjuntos amplos de tópicos retornados pela própria plataforma e selecionava amostras de discussão a partir desse universo, registrando em separado os títulos dos tópicos e a quantidade de comentários coletados por fio. Em seguida, fazia-se a coleta integral das cadeias de comentários disponíveis em cada tópico selecionado, de modo a preservar não apenas enunciados isolados, mas também sequências conversacionais.

O material foi então consolidado em um único arquivo e submetido a limpeza moderada: remoção de URLs, emojis, resíduos da plataforma, identificadores numéricos e outros elementos claramente não textuais; padronização de espaços; normalização de caracteres; e conversão para caixa baixa. Sobre essa base, realizou-se uma segunda camada de regularização lexical, mais extensa do que no corpus dos debates. Essa etapa foi decisiva para reduzir marcas típicas de escrita de plataforma que, se preservadas indiscriminadamente, tenderiam a dominar artificialmente o espaço vetorial.

Entre as intervenções mais importantes dessa segunda camada estiveram: (i) exclusão de marcadores de plataforma e nomes de comunidades; (ii) exclusão de risadas gráficas, interjeições e abreviações interacionais altamente repetitivas; (iii) exclusão de parte do jargão de jogos e de marcas fortemente circunstanciais de performance online; e (iv) regularização de variantes ortográficas, acentuais, abreviações correntes e formas multiword. Essa regularização não teve por objetivo homogeneizar estilisticamente a fala online, mas aproximar variantes evidentes de um mesmo item lexical para que o modelo aprendesse padrões semânticos mais confiáveis.

Distribuição do corpus online por fórum

Fórum	Comentários	% do corpus
Brasil	34.034	22,9
BrasildoB	15.716	10,6
Brasilivre	22.767	15,3
Conversas	28.272	19,0
Desabafos	5.878	4,0
Investimentos	20.458	13,7
SaoPaulo	21.667	14,6
Total	**148.792**	**100,0**

Tabela 5: Distribuição do corpus de comentários online de referência

Intervenções lexicais documentadas

As rotinas preservadas de regularização lexical mostram, com bastante nitidez, que os dois corpora exigiram tratamentos distintos. No corpus dos debates, a intervenção lexical explícita é curta e fortemente orientada por ruído de transmissão: remove um conjunto muito pequeno de resíduos e unifica apenas alguns topônimos recorrentes. No corpus online, a intervenção é mais extensa porque o material traz, de forma muito mais concentrada, marcas de plataforma, abreviação, risada gráfica, jargão circunstancial, ortografia variável e duplicação de variantes semântico-equivalentes.

Essa diferença não deve ser lida como "correção maior" de um corpus e "correção menor" de outro em sentido valorativo. Ela reflete a natureza de cada objeto. O debate transcrito concentra problemas de reconhecimento automático de fala e de resíduos de transmissão; os comentários online concentram problemas de escrita informal, repetição gráfica, siglas de circulação comunitária e variantes ortográficas em competição. Em ambos os casos, a intervenção procurou ser conservadora: remover ruído altamente previsível e aproximar variantes evidentes de uma mesma unidade lexical, sem apagar a heterogeneidade discursiva relevante.

Família de regularização no corpus online	**Função analítica**	**Exemplos ilustrativos**
Marcadores de plataforma e comunidade	impedir que nomes de subfóruns e rotinas da plataforma dominem a vizinhança semântica	*rbrasil*, *rbrasilivre*, *rinvestimentos*, *reddit*
Risada gráfica e ruído interacional	reduzir repetição gráfica sem conteúdo temático robusto	*kkk*, *kkkkk*, *rs*, *rsrs*, *haha*, *hahaha*
Abreviações conversacionais muito circunstanciais	diminuir dispersão causada por marcas interacionais pouco estáveis	*pfv*, *pls*, *vdd*, *oq*, *sqn*, *flw*, *vlw*
Regularização ortográfica e acentual	aproximar variantes evidentes de uma mesma unidade lexical	*politica* → *política*; *educacao* → *educação*; *familia* → *família*
Expansão de abreviações frequentes	concentrar coocorrências hoje dispersas em formas abreviadas	*vc* → *você*; *cmg* → *comigo*; *pra* → *para*
Unificação temática e multiword	preservar entidades e tópicos recorrentes como unidades semânticas estáveis	*rio de janeiro* → *rio_de_janeiro*; *são paulo* → *são_paulo*; *covid/corona* → *coronavírus*

Tabela 6: Famílias de regularização lexical no corpus de comentários online

Nota de comparabilidade e alcance contrastivo

Os dois corpora não são equivalentes em gênero, extensão nem condições de produção, e o suplemento não tenta apagar essa assimetria. O corpus dos debates corresponde a um gênero institucional relativamente estável, com circulação pública regulada, temporalidade concentrada e expectativas explícitas de performance. O corpus online, por sua vez, é mais heterogêneo em tema, extensão e forma de interação. A comparação não tem, portanto, finalidade representacional. Seu valor está em construir um contraste entre campos simbólicos produzidos sob regimes diferentes de enunciação, para observar se conceitos politicamente relevantes são conectados por mediações igualmente distintas. É por essa razão que o segundo corpus funciona como interlocutor contrastivo, e não como retrato estatístico do país.

A.4 Rotinas de avaliação e controle de qualidade

Além da construção e da limpeza dos corpora, o projeto preservou rotinas diagnósticas aplicadas separadamente a cada modelo treinado. Essas rotinas não constituem a evidência principal do artigo e não substituem a interpretação discursiva dos caminhos. Sua função é mais básica e, ao mesmo tempo, decisiva: verificar se o espaço vetorial resultante exibe cobertura lexical, distribuição vetorial e vizinhanças semânticas compatíveis com um uso analítico responsável.

Em termos práticos, essas rotinas examinam cinco aspectos. Primeiro, verificam estatísticas básicas do modelo, como tamanho do vocabulário e dimensionalidade. Segundo, observam propriedades globais dos vetores, especialmente a norma média e sua dispersão. Terceiro, testam relações semânticas simples em pares lexicais gerais, como *bom–ruim*, *feliz–triste* e *rico–pobre*. Quarto, executam analogias fechadas e abertas em campos institucionais, políticos e sociais, não para "provar" conteúdo político do modelo, mas para verificar se o espaço responde de forma semanticamente plausível a relações estruturais relativamente estáveis. Quinto, inspecionam vizinhanças lexicais e sua coerência em domínios como sentimentos, política, vida social, economia e problemas públicos.

Esses testes devem ser lidos como controle de sanidade semântica do espaço distribuicional. Um modelo pode falhar por cobertura insuficiente, por colapso de vizinhanças, por dispersão vetorial anômala ou por incapacidade de recuperar relações básicas entre palavras de alta frequência. A presença dessas rotinas no suplemento tem precisamente a função de mostrar que a pesquisa não saltou da coleta ao resultado interpretativo sem verificar antes a qualidade mínima do material vetorial sobre o qual os caminhos discursivos geométricos seriam calculados.

Frente diagnóstica	**O que observa**	**Função para a pesquisa**
Estatísticas básicas do modelo	tamanho do vocabulário e dimensionalidade	detectar modelos degenerados ou cobertura manifestamente insuficiente
Qualidade geométrica global	norma média dos vetores e dispersão	verificar se o espaço apresenta distribuição vetorial plausível
Relações semânticas simples	similaridade em pares lexicais gerais (*bom–ruim*, *feliz–triste*, *rico–pobre* etc.)	checar se o modelo preserva relações mínimas de proximidade e contraste
Analogias fechadas	relações com resposta esperada em domínios institucionais e sociais	testar plausibilidade estrutural do espaço em relações mais complexas
Analogias abertas	continuidade de cadeias relacionais sem termo-alvo fixado	inspecionar a orientação semântica do espaço sem impor resposta única
Vizinhanças lexicais	vizinhos mais próximos de palavras gerais e políticas	observar densidade local e coerência interpretável das vizinhanças
Coerência de vizinhança	média de similaridade entre vizinhos de um mesmo termo	detectar agrupamentos pouco coesos ou semanticamente frágeis
Cobertura por domínios	presença de palavras-chave em categorias como política, sentimentos, tópicos e problemas públicos	verificar se o vocabulário treinado cobre minimamente os domínios relevantes
Síntese diagnóstica final	composição de métricas exportadas em quadro resumido	registrar, de forma auditável, a avaliação global de cada espaço

Tabela 7: Rotinas diagnósticas preservadas para controle de qualidade dos modelos

Natureza e limite dessas avaliações

As avaliações acima não devem ser confundidas com validação exaustiva do estudo nem com substituto das checagens de robustez dos próprios caminhos discursivos geométricos.

Elas operam em outro nível. Seu papel é mostrar que o treinamento não produziu um espaço semanticamente colapsado antes mesmo de começarmos a perguntar pelos pares-âncora do artigo. Em outras palavras, são rotinas de confiabilidade do suporte vetorial, não da interpretação sociolinguística final.

Também por isso, os resultados dessas avaliações não foram incorporados ao corpo do artigo como seção autônoma. A pergunta central do manuscrito não é "quão bom" é o modelo em termos gerais, mas que tipo de mediação semântica emerge quando ele é mobilizado para comparar dois campos discursivos. Ainda assim, registrar essas rotinas no suplemento é importante porque elas documentam a seriedade do preparo analítico e tornam mais transparente o caminho entre corpus e interpretação.

A.5 Verificações complementares de robustez e critérios de leitura

O artigo principal anuncia três frentes de verificação complementar: estabilidade das rotas em reexecuções, sensibilidade ao tamanho do corpus online e comparação com versões menos restritivas do procedimento. Neste suplemento, o ponto central é explicitar o que conta como robustez para esta pesquisa. Não se trata de exigir identidade literal de todos os intermediários em toda reexecução, mas de perguntar se o *núcleo funcional* da rota se mantém: isto é, se o caminho continua a ser mediado predominantemente por instâncias técnico-administrativas no debate e por eventos, antagonistas, atores nomeáveis ou operações econômicas concretas no corpus online.

Estabilidade entre reexecuções

Em procedimentos de treinamento distribuicional, pequenas variações locais são esperáveis, sobretudo quando não se fixa uma semente global para o treinamento do modelo. Por essa razão, a estabilidade relevante para os caminhos discursivos geométricos não deve ser confundida com repetição literal de cada vocábulo em cada posição. O critério adotado neste projeto é mais compatível com a pergunta linguística do artigo: considera-se estável uma rota cujo eixo de mediação se preserva, ainda que alguns itens marginais se alterem.

Na prática, isso significa distinguir entre *intermediários nucleares* e *intermediários periféricos*. Os primeiros são palavras cuja contribuição interpretativa é forte e coerente com o restante do caminho; os segundos podem variar mais sem comprometer a leitura geral, sobretudo quando são itens opacos, avaliativos ou pouco específicos. Assim, a estabilidade entre reexecuções é julgada por manutenção do perfil discursivo do caminho, e não por identidade

lexical absoluta.

Sensibilidade ao tamanho do corpus online

A diferença de escala entre os corpora exige uma segunda verificação: até que ponto o contraste interpretado depende do volume do corpus online. O ponto decisivo, aqui, não é produzir um espelhamento artificial entre os conjuntos, mas observar se a redução do corpus online afetaria justamente os intermediários que sustentam a leitura de historicização, antagonismo e materialidade social. Em outras palavras, a pergunta não é apenas quantitativa; é funcional. Mesmo quando uma subamostra altere palavras específicas, interessa saber se ela preserva o tipo de mediação predominante.

Essa verificação é particularmente importante porque o corpus online reúne comunidades com temas e registros diversos. Se o padrão observado desaparecesse inteiramente sob pequenas alterações de escala ou composição, o alcance do artigo teria de ser reduzido. Se, ao contrário, o núcleo de mediação se mantiver, a comparação ganha credibilidade mesmo sem pretender equivalência sociológica entre os corpora.

Comparação com versões menos restritivas do procedimento

A terceira verificação diz respeito à utilidade do critério de persistência dimensional. O artigo não introduz a persistência como ornamento formal, mas como mecanismo para reduzir a presença de intermediários ocasionais. Em versões menos restritivas do procedimento — por exemplo, sem persistência dimensional ou com limiar reduzido — a tendência esperável é o aumento de itens localmente próximos ao ponto médio, mas semanticamente menos estáveis no conjunto das projeções. Em termos interpretativos, isso costuma resultar em caminhos mais longos, menos parcimoniosos e mais expostos a ruído.

Por essa razão, o limiar de 30% foi mantido como compromisso entre abertura lexical e legibilidade interpretativa. Em corpora ainda relativamente modestos, limiares muito altos poderiam eliminar excessivamente a mediação; limiares muito baixos, por sua vez, tenderiam a povoar os caminhos com itens cuja contribuição discursiva é fraca ou contingente. O critério adotado, portanto, não garante verdade automática, mas oferece um ponto de equilíbrio metodológico compatível com a pergunta do artigo.

Critério geral de robustez interpretativa

As três frentes acima convergem para um mesmo princípio. A robustez que interessa a este estudo é *discursivo-funcional*: um resultado é mais forte quando diferentes verificações preservam a oposição entre mediação técnico-administrativa no corpus do debate e mediação historicamente ancorada, antagonística ou materialmente concreta no corpus online. Mudanças localizadas em itens opacos não anulam esse contraste; ao contrário, tornam ainda mais importante separar o núcleo interpretável da periferia lexical do caminho.

A.6 Resultados suplementares e critérios de interpretação

Tabela consolidada dos cinco pares focais

Tabela 8: Caminhos discursivos geométricos dos cinco pares focais

Par	Corpus do debate	Corpus online de referência
democracia–ditadura	Democracia → Ambientalista → Comunidade → Justamente → Valorização → Parlamentares → MTST → Mentalidade → Smart Sampa → UBSs → Esquerdista → Comunista → Ditadura	Democracia → Monarquia → Pretexto → Impeachment → Democrática → Liberais → Massacre → República → Massacres → Comunista → Ditadura
dinheiro–elite	Dinheiro → Ministério → Impressionante → Compromisso → História → Jornalistas → Economia → Realmente → Rio de Janeiro → Política → Empresarização → Administração → Elite	Dinheiro → Aluguel → Empréstimo → Bilionário → Tesouro → Financiamento → Clientes → Famílias → Patrimônio → Arrecadado → Financeiramente → Poupança → Desigualdade → Milionários → Elite

Par	Corpus do debate	Corpus online de referência
governo–população	Governo → Educação → Empresas → Questão → Pública → Gestão → Nessa → Trânsito → Ainda → Rachadinha → Janones → Oportunidade → População	Governo → Transformando → Federal → Governadores → Socialista → Polícias → Homicídios → Nordeste → Rio de Janeiro → Liberais → População
pobres–elite	Pobres → Marginais → Esquerdista → Criminalidade → Nordestinos → Trabalhadores → Comunidades → Alfabetização → Solidariedade → Brasileiros → Bandidagem → Cidadãos → Sociedade → Elite	Pobres → Cidadãos → Capitalistas → Intelectuais → Reacionários → Milionários → Dominantes → Monarquistas → Elites → Instância → Elite
população–desemprego	População → Olivato (Eduardo Olivatto) → Democracia → Investimentos → Violência → Incompetência → Desenvolvimento → Especialistas → Missionária → Empreendedorismo → Centro-direita → Praticamente → Desemprego	População → Socialista → Segurança → Reflexo → Homicídios → Diminuindo → Natalidade → Pobreza → Banqueiros → Generalizado → Bacen → Menor → Desemprego

Codificação funcional expandida dos intermediários

A tabela a seguir explicita a camada agregada de evidência apenas resumida no corpo do artigo. Não se trata de uma taxonomia fechada, mas de uma codificação funcional mínima para tornar mais visível a recorrência de famílias de mediação entre os dois corpora.

Tabela 9: Codificação funcional de intermediários selecionados

Par	**Corpus**	**Intermediários de maior peso interpretativo**	**Categoria funcional predominante**	**Observação**
democracia–ditadura	Debate	Parlamentares; MTST; Smart Sampa; UBSs; Esquerdista; Comunista	mediação político-institucional e gestão urbana	conflito reencaminhado por léxico administrativo e institucional
democracia–ditadura	Online	Monarquia; Impeachment; Massacre; República; Massacres; Comunista	memória histórica, ruptura política e trauma coletivo	a passagem chega à ditadura por eventos e formas políticas reconhecíveis
dinheiro–elite	Debate	Ministério; Jornalistas; Política; Empresarização; Administração	legitimação institucional e enquadramento administrativo	a elite aparece mediada por instâncias formais e linguagem pública
dinheiro–elite	Online	Aluguel; Empréstimo; Financiamento; Patrimônio; Poupança; Desigualdade; Milionários	economia cotidiana, acumulação e desigualdade	o caminho sobe do cotidiano financeiro à estrutura desigual de riqueza
governo–população	Debate	Educação; Questão Pública; Gestão; Trânsito; Oportunidade	inventário de gestão pública e administração de temas	a população emerge como destinatária de uma pauta administrável

Par	Corpus	Intermediários de maior peso interpretativo	Categoria funcional predominante	Observação
governo–população	Online	Federal; Governadores; Polícias; Homicídios; Nordeste; Rio de Janeiro	poder estatal concreto, territorialização e violência	a mediação passa por aparato público, conflito e efeitos sociais tangíveis
pobres–elite	Debate	Marginais; Criminalidade; Nordestinos; Trabalhadores; Alfabetização; Solidariedade; Cidadãos; Sociedade	integração social burocratizada com criminalização intermediária	o trajeto é longo e internamente tenso: integra e criminaliza ao mesmo tempo
pobres–elite	Online	Capitalistas; Reacionários; Milionários; Dominantes; Monarquistas; Elites	antagonismo de classe e nomeação direta de dominantes	o conflito é nomeado com muito menos amortecimento institucional
população–desemprego	Debate	Investimentos; Violência; Incompetência; Desenvolvimento; Especialistas; Empreendedorismo; Centro-direita	causalidade gerencial e solução ideológica em chave administrativa	o desemprego aparece mediado por repertório de gestão e especialização

Par	Corpus	Intermediários de maior peso interpretativo	Categoria funcional predominante	Observação
população–desemprego	Online	Segurança; Homicídios; Natalidade; Pobreza; Banqueiros; Bacen	indicadores sociais concretos e atores econômicos nomeáveis	a rota cruza efeitos sociais e instâncias econômicas materializadas

Itens opacos, raros ou idiossincráticos

Alguns intermediários exigem cautela interpretativa especial. No par *democracia–ditadura*, itens como *Ambientalista*, *Justamente* e *Valorização* são menos transparentes do que *Parlamentares*, *Smart Sampa* ou *Massacre*. Em *dinheiro–elite*, *Impressionante* e *Realmente* têm peso interpretativo menor do que *Ministério*, *Administração* ou *Desigualdade*. Em *governo–população*, *Nessa*, *Ainda* e *Transformando* são itens de leitura fraca se considerados isoladamente, ao passo que *Gestão*, *Polícias* e *Homicídios* organizam efetivamente a rota.

No par *pobres–elite*, *Instância* no corpus online é o elo menos nítido do caminho; ainda assim, ele não neutraliza a cadeia fortemente antagonística que o antecede. Já em *população–desemprego*, *Olivato*, *Missionária*, *Praticamente*, *Reflexo*, *Generalizado* e *Menor* merecem ser tratados como itens periféricos ou contextualmente opacos. O ponto metodológico importante é que esses vocábulos não foram eliminados por edição interpretativa. Eles permanecem registrados para que o leitor veja a textura real dos caminhos e compreenda que a análise repousa na convergência dos intermediários mais densos, não na transparência uniforme de todos os elos.

Síntese suplementar do contraste

Tomados em conjunto, os cinco pares focalizados reforçam a leitura apresentada no corpo do artigo. No corpus do debate, os caminhos tendem a atravessar instâncias de administração, institucionalidade, gestão e legitimação burocrática, ainda quando abordam conceitos politicamente carregados. No corpus online de referência, as mediações tendem a acionar memória histórica, eventos traumáticos, antagonistas nomeáveis, efeitos sociais materializados e circuitos de economia cotidiana. O valor do suplemento é tornar mais visível que essa

formulação não depende de um único exemplo nem de uma leitura indiferente à opacidade de alguns itens. Ela se apoia numa recorrência funcional observável ao longo dos cinco caminhos e numa disciplina interpretativa que distingue núcleo e periferia lexical.

O que ficou fora do escopo analítico

O desenho do estudo deixou de fora várias perguntas potencialmente relevantes. Não se realizou modelagem separada por candidato, nem comparação entre turnos do debate, nem segmentação temporal fina da campanha. Também não se buscou expandir o conjunto de pares-âncora para cobertura semântica mais ampla, nem comparar os resultados com outros gêneros institucionais. Por fim, o artigo não pretende derivar, a partir deste corpus online, inferências sociológicas fortes sobre a população brasileira. Essas exclusões não são falhas acidentais, mas escolhas de escopo: o objetivo foi testar, em um estudo de caso contrastivo, se a métrica dos caminhos discursivos geométricos é capaz de tornar visíveis diferenças de mediação semântica teoricamente interpretáveis.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.