

Estado da publicação: O preprint não foi publicado em outro meio.

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: 2026-10-02 (versão 4)

(AAAA-MM-DD)

Justificativa da versão: Versão final do manuscrito pré-decisão editorial.

Tipo de Trabalho: Perspectiva Crítica

Caminhos Discursivos Geométricos e a Geometria do Discurso: o que se oculta nas mediações semânticas

Alexandre Menezes Barroso

(<https://orcid.org/0009-0005-2794-767X>)¹

Terceira versão, outubro 2026. Manuscrito submetido para revisão por pares.

Resumo

Este artigo defende que, em contextos eleitorais marcados por polarização, plataformização e disputa sobre desinformação, a Linguística precisa observar não apenas o que o discurso declara, mas também o que ele torna indireto, administrável ou pouco visível entre conceitos. A partir dessa posição, propõe-se a noção de *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. Os caminhos não são tratados como ponto de chegada da argumentação, mas como ponto de partida para uma discussão mais ampla sobre mediação semântica, tecnocratização e ocultação discursiva. O texto mobiliza, como caso de apoio, os debates eleitorais paulistanos de 2024 e um corpus de comentários online de referência, analisados em espaços treinados separadamente. Nos cinco pares e nas configurações examinadas, a leitura sugere que o corpus do debate institucional tende a amortecer passagens sensíveis por meio de vocabulário técnico-administrativo ou institucional, enquanto o corpus online tende a acionar eventos, antagonistas, agentes financeiros e cadeias da economia cotidiana. O argumento central é que aquilo que parece apenas diferença lexical pode revelar modos distintos de tornar conflitos mais ou menos nomeáveis. A contribuição do artigo é conceitual e metodológica: mostrar como uma métrica vetorial pode servir à interpretação textual sem substituir a análise discursiva, e como a política aparece não só nas palavras escolhidas, mas nos trajetos que ligam uma palavra à outra.

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

¹Doutorando pela Universidade Estadual de Campinas (UNICAMP), no Instituto de Estudos da Linguagem (IEL), localizado em Campinas, São Paulo. Contato: abarroso@outlook.com.br.

Geometric Discursive Paths and the Geometry of Discourse: What Is Hidden in Semantic Mediations

Alexandre Menezes Barroso

Abstract: This article argues that, in electoral contexts marked by polarization, platformization, and disputes over disinformation, Linguistics must examine not only what discourse states, but also what it makes indirect, administrable, or less visible between concepts. From this position, it proposes the notion of *geometric discursive paths*: sequences of intermediate words connecting two anchor concepts in a vector space and filtered by a dimensional-persistence criterion. These paths are not treated as the endpoint of the argument, but as a starting point for a broader discussion of semantic mediation, technocratization, and discursive concealment. The article uses, as supporting evidence, the 2024 São Paulo mayoral debates and an online comments reference corpus, analyzed in separately trained spaces. Within the five concept pairs and model configurations examined, the reading suggests that the institutional debate corpus tends to cushion sensitive transitions through technical-administrative or institutional vocabulary, whereas the online corpus tends to activate events, antagonists, financial actors, and everyday economic chains. The central argument is that what may seem to be mere lexical difference can reveal different ways of making conflicts more or less nameable. The article's contribution is conceptual and methodological: it shows how a vector metric may support textual interpretation without replacing discourse analysis, and how politics appears not only in selected words, but also in the paths that connect one word to another.

Keywords: text analysis; text linguistics; political discourse; vector semantics; semantic mediation.

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, democracia, ditadura, relação entre esferas de poder e disputa sobre instituições apareceram como temas de campanha e de definição pública dos problemas da cidade. Nesse cenário, o debate eleitoral interessa menos por *o que* diz isoladamente do que por *como* organiza passagens entre conceitos politicamente densos.

A tese deste artigo é que parte decisiva do funcionamento político do discurso está naquilo que fica entre os conceitos. O discurso não oculta apenas quando silencia. Ele também oculta quando interpõe mediações aparentemente neutras, técnicas, administrativas ou morais entre termos que poderiam ser ligados de maneira mais direta. Se a passagem entre *população* e *desemprego* atravessa *empreendedorismo*, *investimento* e *oportunidade*, temos um tipo de construção. Se atravessa *pobreza*, *banqueiros* e *segurança*, temos outro. O problema linguístico não é apenas que palavras aparecem, mas o trajeto que torna uma associação mais fácil, mais aceitável ou mais difícil de enunciar.

Essa pergunta se torna ainda mais atual no ciclo eleitoral brasileiro de 2026. A disputa nacional ocorre depois das eleições municipais de 2024 e em um ambiente no qual a própria Justiça Eleitoral já trata inteligência artificial, deepfakes, conteúdos sintéticos e desinformação como temas centrais da campanha (TRIBUNAL SUPERIOR ELEITORAL, 2026a; TRIBUNAL SUPERIOR ELEITORAL, 2026b). Essa moldura pública é importante, mas não basta. O risco linguístico do período eleitoral não está apenas no conteúdo falso, manipulado ou fabricado. Está também na forma mais comum, menos espetacular e mais difícil de detectar pela checagem factual: a reorganização das mediações semânticas que fazem um conflito parecer gestão, uma desigualdade parecer oportunidade, uma violência parecer segurança ou uma crise democrática parecer apenas problema de comunicação.

É nesse ponto que os *caminhos discursivos geométricos* entram no argumento. A expressão designa sequências de palavras intermediárias que conectam dois conceitos-âncora em um espaço vetorial e são filtradas por um critério de persistência dimensional. A métrica é importante, mas não é o centro do texto. Ela funciona como instrumento de localização de indícios: destaca elos entre conceitos e, com isso, permite perguntar que tipo de mediação um campo discursivo naturaliza. Os caminhos são, portanto, o ponto de partida para uma discussão sobre o que o discurso torna visível e o que ele torna menos nomeável.

Essa posição se inscreve em uma discussão mais ampla sobre comunicação política,

circulação digital e conflito discursivo. Estudos recentes mostram que as mídias digitais remodelam a comunicação política brasileira, intensificam a circulação de mensagens segmentadas e favorecem formas de disputa em que linguagem, afeto e antagonismo se sobrepõem (ITU-ASSU et al., 2023; TUCCI; GOUVEIA, 2025; OLIVEIRA; MIRANDA, 2024; OLIVEIRA et al., 2025). 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. A pergunta, então, não é apenas quais temas aparecem na eleição, mas como certos temas são ligados uns aos outros de modo a proteger, deslocar ou intensificar conflitos.

Para sustentar essa discussão, tomo como caso de apoio 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 segundo corpus não representa o Brasil, nem funciona como sinônimo de opinião pública. Ele é um interlocutor contrastivo. Sua função é permitir que se observe se, fora do enquadramento institucional do debate, os mesmos conceitos são conectados por mediações semelhantes ou distintas.

Os cinco pares analisados são *democracia–ditadura*, *dinheiro–elite*, *governo–população*, *pobres–elite* e *população–desemprego*. Eles foram escolhidos por condensarem problemas centrais para a interface entre Linguística, política e vida social. Não esgotam o campo semântico dos corpora, nem pretendem demonstrar uma teoria total do discurso político brasileiro. Funcionam como cenas de leitura. A partir delas, argumento que, nesses casos e nessas configurações, o corpus do debate institucional tende a amortecer passagens sensíveis por vocabulário técnico-administrativo, institucional ou moral-gerencial, enquanto o corpus online tende a ativar eventos, antagonistas nomeados, agentes financeiros, violência, território e cadeias da economia cotidiana.

A relevância dos caminhos está justamente nesse deslocamento. Eles não substituem a interpretação; eles indicam onde a interpretação precisa olhar. Em vez de perguntar apenas se um discurso é falso, agressivo, populista, tecnocrático ou polarizado, os caminhos permitem perguntar: quais palavras estão fazendo a ponte? Que elos tornam o conflito menos visível? Que intermediários fazem a política aparecer como gestão, a desigualdade como economia, a violência como segurança ou a crise democrática como ruído comunicacional? Essa é a contribuição que este artigo busca oferecer à Linguística textual e às suas interfaces.

O artigo organiza-se assim: a Seção 2 apresenta a fundamentação teórica; a 3, o procedimento de apoio e a transparência analítica; a 4, o material de contraste; a 5, os cinco caminhos como ponto de partida interpretativo; a 6, a discussão sobre ocultação, mediação e atualidade eleitoral; e a 7, a conclusão.

2 Fundamentação teórica: mediação, ocultação e texto como espaço relacional

O ponto de partida teórico deste artigo 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 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.

A palavra decisiva aqui é *mediação*. O sentido político de um conceito raramente aparece apenas na palavra tomada isoladamente. Ele aparece também na vizinhança que a acompanha, nos intermediários que a aproximam de outra palavra e nos desvios que tornam uma conexão mais longa, mais indireta ou mais administrável. Nessa perspectiva, ocultar não significa apenas não dizer. Pode significar dizer por meio de uma cadeia que desloca o conflito para uma zona menos conflitiva. Uma relação social pode continuar presente, mas recoberta por vocabulário de gestão, mérito, oportunidade, investimento, ordem ou segurança.

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 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. A passagem entre esses níveis não é uma equivalência: a proximidade vetorial resume semelhanças de contextos, e a posição intermediária seleciona candidatos a articular repertórios associados aos dois âncoras. A persistência indica estabilidade relativa às projeções aninhadas de uma mesma base, não confirmação por observações independentes nem significância estatística. A mediação discursiva é a hipótese interpretativa formulada quando esses candidatos compõem relações pertinentes ao problema e às condições de produção examinadas. Palavras próximas podem também compartilhar um tópico, expressar oposição ou reunir sentidos distintos em um único vetor. Por isso, a rota não reproduz necessariamente uma sequência enunciada, não demonstra intenção de ocultar e não mede efeitos sobre o leitor. A noção de campo simbólico autoriza investigar relações; não autoriza identificar automaticamente posição geométrica e função discursiva.

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 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: 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 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 surgem intermediários ligados a personagens políticos, antagonismos sociais, instituições financeiras, violência 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 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. O ponto comum é este: aquilo que se oculta no discurso pode estar menos na ausência de uma palavra do que no caminho que torna uma relação menos direta.

3 Como os caminhos entram no argumento

Esta seção documenta o procedimento que sustenta a evidência mobilizada pelo argumento. Sua função não é deslocar o centro do texto para o reporte de resultados de corpus, mas mostrar de modo auditável como os caminhos foram obtidos, quais decisões interpretativas foram tomadas e por que os dados podem funcionar como apoio a uma tese sobre mediação semântica. Indicamos o que foi feito, em coerência com a “anotação linguística de corpora” (PEREIRA E SILVA, 2024, p. 2369), remetendo estatísticas, critérios de verificação e casos-limite ao suplemento. Essa economia expositiva é compatível com a construção de um “pipeline de validação seguro e confiável” (RUYTERS et al., 2026, p. 28, tradução nossa)², quando o artigo resume a arquitetura analítica e desloca detalhes operacionais ao apoio suplementar.

Construção dos espaços vetoriais

Cada corpus gera um espaço vetorial próprio. Em vez de projetar debate e comentários num único modelo, partimos de vetores fastText (GRAVE et al., 2018) produzidos pela equipe do Facebook AI Research para o português, com dados do Common Crawl e da Wikipédia, 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)³. 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. O recurso inicial é o arquivo `cc.pt.300.vec`, referente ao português, sem restrição exclusiva à variedade brasileira. Seu pré-treinamento empregou CBOW com pesos posicionais, 300 dimensões, n-gramas de cinco caracteres, janela 5 e dez exemplos negativos. Transferimos vetores lexicais para Word2Vec, não o modelo sublexical completo do fastText.

Em ambos os casos, os vetores têm 300 dimensões e partem do mesmo recurso inicial

²“secure and reliable validation pipeline” (RUYTERS et al., 2026, p. 28)

³“using their methodology of directly training on all the available data was not feasible in our case” (FREDÉN et al., 2024, p. 9)

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., 2026, p. 30, tradução nossa)⁴ 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 (MIKOLOV et al., 2013): 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 limita a comparação: arquitetura, janela e frequência mínima afetam a geometria interna, a seleção em torno do ponto médio e a ordenação dos intermediários. O contraste descreve os dois corpora sob essas configurações, mas não permite atribuir as diferenças exclusivamente ao gênero discursivo. A comparação de rotas, em vez de distâncias entre espaços, não elimina esse confundimento. 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)⁵, 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)⁶.

Os gráficos bidimensionais têm função apenas ilustrativa. Os cálculos dos caminhos discursivos geométricos partem do espaço de 300 dimensões e de suas projeções sucessivas, não das 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., 2025, p. 5, tradução nossa)⁷.

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

⁴“transparency and explainability” (RUYTERS et al., 2026, p. 30)

⁵“simplify ideological positions into a binary spectrum” (CHEN et al., 2024, p. 3)

⁶“potential risks brought by uncertain annotation” (CHEN et al., 2024, p. 3)

⁷“move beyond surface-level linguistic features and engage with the broader contextual dimensions of language” (KRYKONIUK et al., 2025, p. 5)

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 por distância euclidiana. Entre eles, excluem-se os próprios âncoras e retêm-se os itens cuja projeção escalar no eixo orientado do primeiro para o segundo âncora caia entre as extremidades. O critério não exige que o vetor esteja sobre o segmento: limita sua posição longitudinal, enquanto a distância ao ponto médio controla a seleção inicial. Os comprimentos envolvidos são calculados para cada par e dimensão, não fixados como parâmetros; as expressões constam do Apêndice, em “Extração dos caminhos discursivos geométricos”.

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; operacionalmente, exigem-se mais de 90 ocorrências, isto é, ao menos 91 nas 299 projeções de 300 a 2 dimensões. Esse limiar é heurístico: um valor mais alto pode produzir caminhos mais curtos, enquanto um valor mais baixo admite mais candidatos. O interesse é distinguir mediações ocasionais de padrões recorrentes e permitir um levantamento quantitativo dessas ocorrências.

Quarto, as palavras robustas são ordenadas pela média de suas projeções escalares, medidas a partir do primeiro âncora na direção do segundo, em 300, 250, 200, 150 e 100 dimensões. A ordem é crescente e não corresponde à distância ao ponto médio; este é usado apenas na seleção inicial. A média utiliza os comprimentos projetados, sem normalização pela distância entre os âncoras. 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 examinar regularidades em múltiplas projeções do mesmo espaço, sem equiparar persistência geométrica a validade discursiva, aproximando-se da concepção de que o espaço discursivo funciona como “um sistema dinâmico de estados discursivos possíveis” (KRYKONIUK et al., 2025, p. 6, tradução nossa)⁸.

⁸“a dynamic system of possible discourse states” (KRYKONIUK et al., 2025, p. 6)

Especificação analítica resumida

Item	Especificação resumida
Recurso inicial	Vetores fastText para o português (Common Crawl e Wikipédia).
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 de ajuste	10, após cinco épocas preliminares de Word2Vec.
Amostragem negativa	15.
Taxa de aprendizado	Inicial 0.025; mínima 0.0001.
Base de projeção	PCA por SVD completa sobre o espaço centralizado; semente 42 no treinamento.
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	Projeção escalar entre os âncoras; exclusão dos próprios âncoras.
Regra de persistência	Ao menos 91 ocorrências nas 299 projeções.
Ordenação final	Projeção escalar média a partir do primeiro âncora em 300, 250, 200, 150 e 100 dimensões.
Pares-âncora do artigo	democracia–ditadura; dinheiro–elite; governo–população; pobres–elite; população–desemprego.
Detalhamento adicional	Estatísticas de corpus, pré-processamento, critérios 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. 7, tradução nossa)⁹, o ganho analítico aqui está em outra pergunta: como diferentes corpora organizam mediações semânticas robustas entre conceitos-âncora.

⁹“outperform benchmark word-level methodology on measuring position similarities in political Twitter datasets” (CHEN et al., 2024, p. 7)

Robustez, escopo e papel da evidência

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)¹⁰. Também por isso, buscamos uma validação interpretável, interessada em “não apenas classificar os enunciados” (PEREIRA E SILVA, 2024, 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)¹¹, 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., 2025, p. 12, tradução nossa)¹² a partir de trajetórias informativas.

Por isso, delimitamos três frentes para ampliar a validação: estabilidade das rotas em reexecuções, sensibilidade ao tamanho do corpus de comentários e comparação com versões menos restritivas do procedimento. O suplemento explicita critérios para essas verificações; a evidência aqui mobilizada consiste nos cinco caminhos preservados e em sua persistência no interior de cada espaço, não em resultados de reamostragem ou de comparação controlada entre arquiteturas. Nessa chave, as checagens não entram como apêndice decorativo, mas como etapas “não como complementares, mas como necessárias” (RUYTERS et al., 2026, p. 45, tradução nossa)¹³ para testar a dependência dos caminhos em relação à configuração adotada. No corpo do artigo, resumimos o necessário; no suplemento, apresentamos os caminhos e os critérios de desempate, exclusão e verificação para que outros possam “generalizar, complementar ou refutar esses dados” (PEREIRA E SILVA, 2024, p. 2393), em linha com a transparência e reprodutibilidade de outros estudos e pesquisas similares.

O argumento deve ser lido em escala adequada. O que observamos é uma tendência contrastiva recorrente nos pares e nas configurações analisados, com a cautela adicional de que “os termos ideológicos apresentam menor estabilidade” (FREDÉN et al., 2024, p. 15, tradução nossa)¹⁴, 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

¹⁰“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)

¹¹“coincide with social common sense, political news, as well as prior research” (CHEN et al., 2024, p. 7)

¹²“the interpretation and comparison of meaning across texts and registers” (KRYKONIUK et al., 2025, p. 12)

¹³“not as complementary, but as necessary” (RUYTERS et al., 2026, p. 45)

¹⁴“the ideological terms show less stability” (FREDÉN et al., 2024, p. 15)

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 Material de apoio e ponte entre 2024 e 2026

A argumentação mobiliza dois espaços discursivos construídos a partir de corpora distintos em gênero, extensão e condições de produção. A função desta seção é delimitar o material usado como contraste e explicitar 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. Os dois conjuntos não são tratados como amostras sociológicas equivalentes, mas como campos simbólicos produzidos em regimes diferentes de enunciação.

O corpus de 2024 é importante também por seu lugar histórico. Ele pertence ao ciclo eleitoral anterior, mas ajuda a ler o ciclo atual justamente porque os problemas ali observados não desapareceram: a disputa entre gestão, desigualdade, segurança, democracia, instituições e vida econômica reaparece em escala nacional quando o país se aproxima das Eleições de 2026. Assim, os debates paulistanos não são tomados como retrato do Brasil inteiro, mas como arquivo situado de uma forma de dizer a política que permanece relevante.

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. 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 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 de até 1.000 tópicos por comunidade retornados pela ordenação *hot*. Foram reunidos os comentários disponibilizados pela API, sem expansão dos blocos adicionais de comentários; trata-se, portanto, de uma coleta condicionada à seleção e à disponibilidade da plataforma, não de cadeias necessariamente integrais. 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. Os arquivos consolidados contêm 121.732 unidades delimitadas por espaços nos debates e 5.015.065 nos comentários, antes da tokenização para modelagem. Essa diferença de cerca de 41 vezes pode alterar a cobertura e a estabilidade das representações, especialmente de itens raros. A diversidade temática dos fóruns também amplia os repertórios candidatos, enquanto a concentração temática do debate favorece recorrências institucionais. Esses efeitos se somam às diferenças de arquitetura, janela e limpeza lexical: os caminhos não isolam um efeito de gênero nem autorizam equiparar extensão da rota e intensidade de ocultação.

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. 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, 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. O valor heurístico do segundo corpus está menos em representar uma totalidade social do que em oferecer um termo de contraste relativamente heterogêneo fora do enquadramento institucional do debate.

5 Cinco caminhos como ponto de partida

Os cinco caminhos a seguir são extremamente relevantes para o argumento, mas não são o seu ponto final. Eles são a porta de entrada para uma discussão mais ampla: o que se oculta quando conceitos politicamente carregados são ligados por mediações aparentemente neutras, técnicas, morais ou administrativas? Os pares de conceitos-âncora efetivamente fixados na análise são “democracia”–“ditadura”, “dinheiro”–“elite”, “governo”–“população”, “pobres”–“elite” e “população”–“desemprego”. Eles foram escolhidos autoralmente por sua centralidade político-discursiva e por cobrirem domínios distintos do corpus. São tratados como *casos focais*: não esgotam o espaço semântico dos corpora, mas permitem observar como dois espaços vetoriais conectam conceitos de alta carga política por mediações diferentes. A força da seção se apoia na recorrência das famílias de intermediários que reaparecem ao longo dos casos, na distinção entre itens fortemente interpretáveis e itens mais opacos e, nas figuras, na leitura combinada entre visão macroscópica do caminho e detalhe ampliado da zona densa. Critérios adicionais de estabilidade e sensibilidade são explicitados no 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 → Dita- dura	serviços públicos, regulação, comunicação, criminalidade e fechamento ideológico	impeachment, acusação de autoritarismo, golpe, massacre e memória de regime	o debate traduz a crise democrática em chave institucional-comunicacional; o corpus online a dramatiza como ruptura política
Dinheiro → Elite	população, política, compromisso público, economia e empresarização	crédito, financiamento, tesouro, patrimônio familiar, banqueiros e milionários	a elite é mediada por uma gramática econômico-institucional no debate e por acumulação concreta no corpus online
Governo → População	empresas, cidadania, projetos, educação, gestão, investimento e importância pública	federalismo, governadores, segurança, polícias, homicídios e território	a população surge como destinatária de política pública no debate e como corpo exposto a conflito estatal no corpus online
Pobres → Elite	estigma inicial, comunidade, educação, humanidade, solidariedade, cidadania e sociedade	antagonismo de classe, reação política, dominação cultural, milionários e elites dirigentes	o debate integra e ordena; o corpus online nomeia hierarquias e adversários sociais
População → Desemprego	valorização, desenvolvimento, democracia, investimentos, empreendedorismo, oportunidades, trabalhador e serviço público	segurança nacional, homicídios, natalidade, comparação econômica, pobreza e banqueiros	o debate moraliza e gerencializa a empregabilidade; o corpus online associa desemprego a estrutura social e financeira

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) serviços públicos, regulação e gestão de políticas (“ubss”, “arsesp”, “projetos”, “educação”, “gestão”, “investimento”); (ii) vocabulário de tradução institucional da economia e da ação pública (“população”, “política”, “compromisso”, “fundamental”, “economia”, “empresarização”); e (iii) marcadores ideológicos ou moralizantes enquadrados como problemas de condução e solução (“comunista”, “esquerdista”, “centrodireita”, “empreendedorismo”, “incompetência”, “oportunidades”). Já o corpus online recorre sobretudo a: (i) eventos e acusações de ruptura democrática (“impeachment”, “ditatorial”, “golpista”, “massacre”, “república”); (ii) atores e antagonistas concretos (“banqueiros”, “milionários”, “capitalistas”, “reacionários”, “dominantes”, “governadores”); (iii) operações econômicas de crédito, Estado e patrimônio (“empréstimo”, “tesouro”, “financiamento”, “benefício”, “patrimônio”, “famílias”); e (iv) efeitos sociais materializados em

segurança, território e pobreza (“polícias”, “homicídios”, “nordeste”, “rio_de_janeiro”, “natalidade”, “pobreza”). Não se trata de categorias exaustivas nem de classificação independente da hipótese, mas de uma codificação funcional que torna explícita a leitura de conjunto dos cinco casos.

5.1 “Democracia” → “Ditadura”

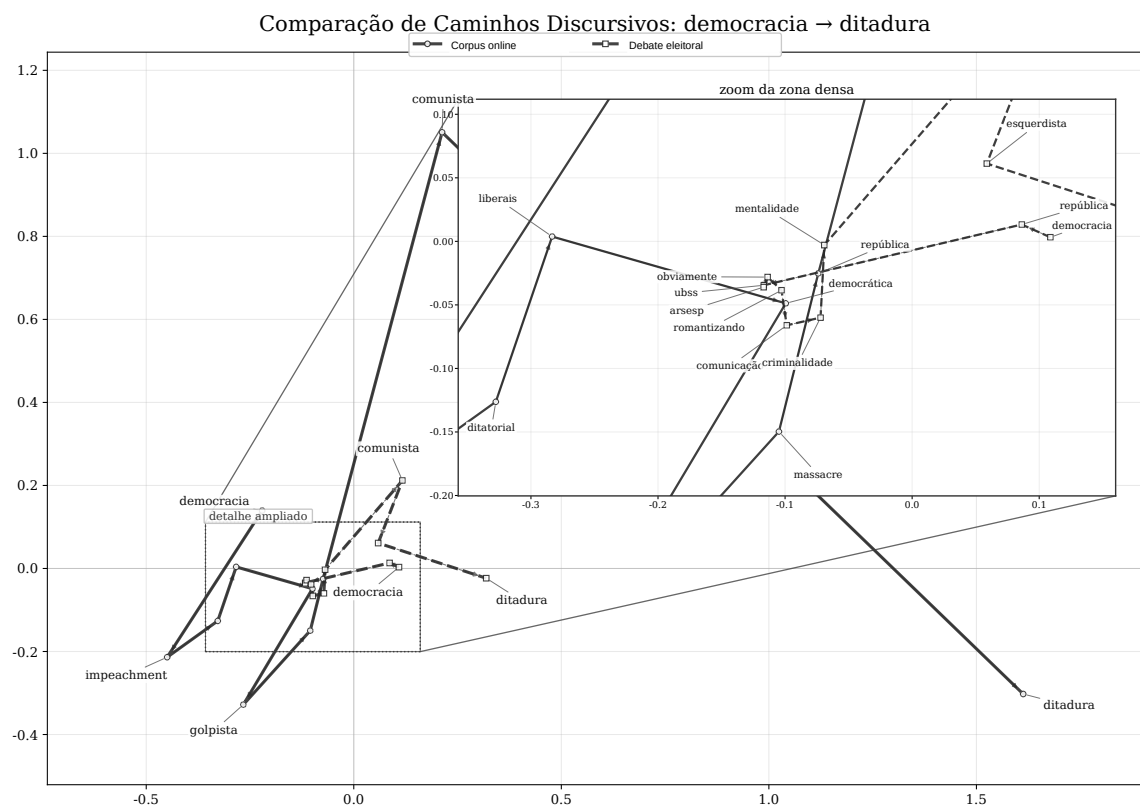


Figura 1: “Democracia” → “Ditadura”, gráfico comparativo entre corpora, com detalhe ampliado da zona densa.

ALT: Gráfico de dispersão comparando os caminhos discursivos para o par democracia–ditadura em dois corpora. O corpus do debate apresenta intermediários ligados a gestão institucional e comunicação; o corpus online apresenta intermediários ligados a ruptura política, impeachment e memória histórica.

Debate	democracia → república → ubss → arsesp → obviamente → romantizando → comunicação → criminalidade → mentalidade → comunista → esquerdista → ditadura
Online	democracia → impeachment → ditatorial → liberais → democrática → golpista → massacre → república → comunista → ditadura

A primeira cena mostra o tipo de ponte que o procedimento torna visível. No corpus do debate, a transição entre “democracia” e “ditadura” começa em “república”, mas logo passa por “ubss” e “arsesp”, isto é, por uma zona de serviços públicos, regulação e institucionalidade cotidiana. A seguir, o caminho deixa de ser propriamente histórico e assume uma textura meta-discursiva: “obviamente”, “romantizando”, “comunicação” e “mentalidade” fazem a passagem entre regime político e avaliação do modo como o adversário fala, enquadra ou imagina a política. A ditadura, nesse percurso, não é alcançada por uma narrativa direta de ruptura institucional; ela emerge depois de uma cadeia que combina Estado-serviço, regulação, comunicação, criminalidade e psicologização do outro. Só ao final aparecem “comunista” e “esquerdista”, como se a oposição democracia–ditadura fosse encerrada por uma classificação ideológica depois de ter sido filtrada por dispositivos de administração e linguagem.

No corpus online, o caminho é mais curto e mais explosivo. “democracia” se aproxima imediatamente de “impeachment”, depois de “ditatorial”, “golpista” e “massacre”. A cadeia faz do conflito democrático um roteiro de acusação: há procedimento institucional, há qualificação autoritária, há disputa sobre quem golpeia quem e há memória de violência. “liberais” e “democrática” não estabilizam o centro do caminho; ao contrário, parecem intensificar sua ambivalência, pois entram no mesmo espaço em que aparecem “ditatorial” e “golpista”. A passagem por “república” antes de “comunista” e “ditadura” fecha o circuito em termos de forma de regime e antagonismo ideológico. O caso, portanto, não mostra apenas dois vocabulários distintos. Mostra duas maneiras de metabolizar a crise democrática: no debate, ela é reencaminhada para serviços, regulação e comunicação; no corpus online, ela é condensada como impeachment, golpe, massacre e disputa de regime.

5.2 “Dinheiro” → “Elite”

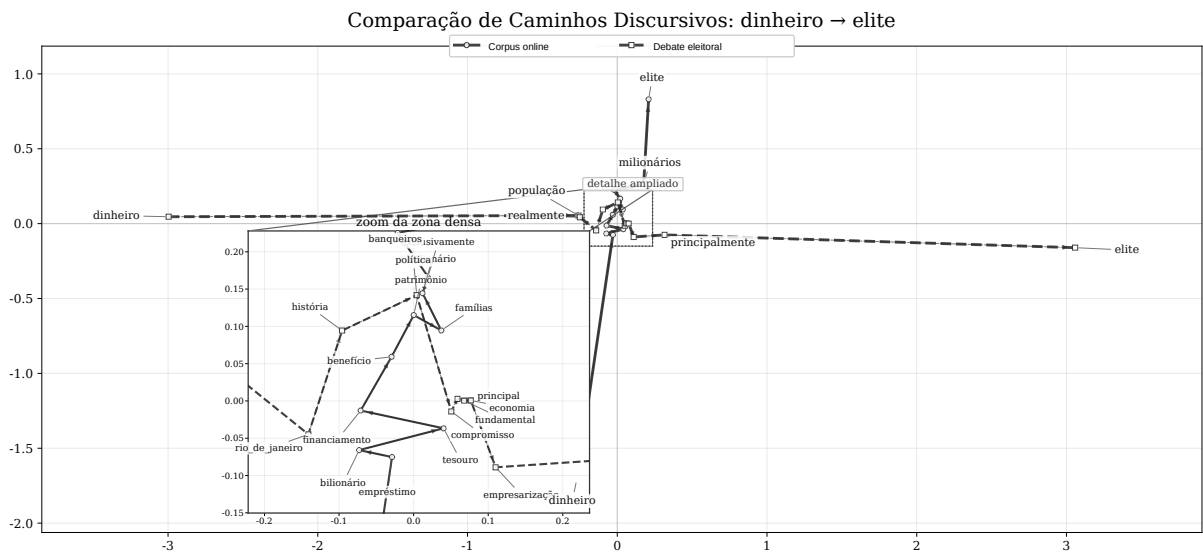


Figura 2: “Dinheiro” → “Elite”, gráfico comparativo entre corpora, com detalhe ampliado da zona densa.

ALT: Gráfico de dispersão comparando os caminhos discursivos para o par dinheiro–elite em dois corpora. O corpus do debate apresenta intermediários ligados à população, política, compromisso público, economia e empresarização; o corpus online apresenta intermediários ligados a empréstimo, financiamento, patrimônio, famílias, banqueiros e milionários.

<i>Debate</i>	dinheiro → população → realmente → rio_de_janeiro → história → política → compromisso → principal → fundamental → economia → empresarização → principalmente → elite
<i>Online</i>	dinheiro → empréstimo → bilionário → tesouro → financiamento → benefício → patrimônio → famílias → milionário → exclusivamente → banqueiros → milionários → elite

Entre “dinheiro” e “elite”, o corpus do debate produz uma rota de nacionalização e formalização da economia. O dinheiro passa por “população”, “rio_de_janeiro”, “história”, “política”, “compromisso”, “principal” e “fundamental” antes de chegar a “economia”, “empresarização” e “elite”. A elite não aparece, portanto, como produto imediato de acumulação ou privilégio financeiro; ela é alcançada por uma sequência que traduz o dinheiro para o plano da

cidade, da história política, da responsabilidade pública e da organização econômica. “empresa-rização” é o elo decisivo, porque desloca o dinheiro para uma racionalidade de gestão e mercado, permitindo que “elite” apareça como desfecho de uma economia institucionalmente narrada, e não como antagonista diretamente nomeado.

No corpus online, em contraste, a cadeia é quase contábil em sua materialidade. “dinheiro” passa por “empréstimo”, “tesouro”, “financiamento”, “benefício”, “patrimônio” e “famílias”; em seguida, a escala muda para “milionário”, “banqueiros”, “milionários” e “elite”. O percurso organiza uma gramática de acumulação: crédito, recursos públicos, financiamento, benefício, patrimônio familiar e agentes financeiros. “bilionário” logo no início e “milionários” no fim comprimem escalas de riqueza que, na experiência comum, aparecem como distantes mas causalmente ligadas. O que localmente se observa é uma mudança de escala: o debate converte dinheiro em política econômica e compromisso institucional; o corpus online converte dinheiro em dívida, patrimônio, bancos e concentração. Onde o debate abstrai e formaliza, o corpus online concretiza e acusa.

5.3 “Governo” → “População”

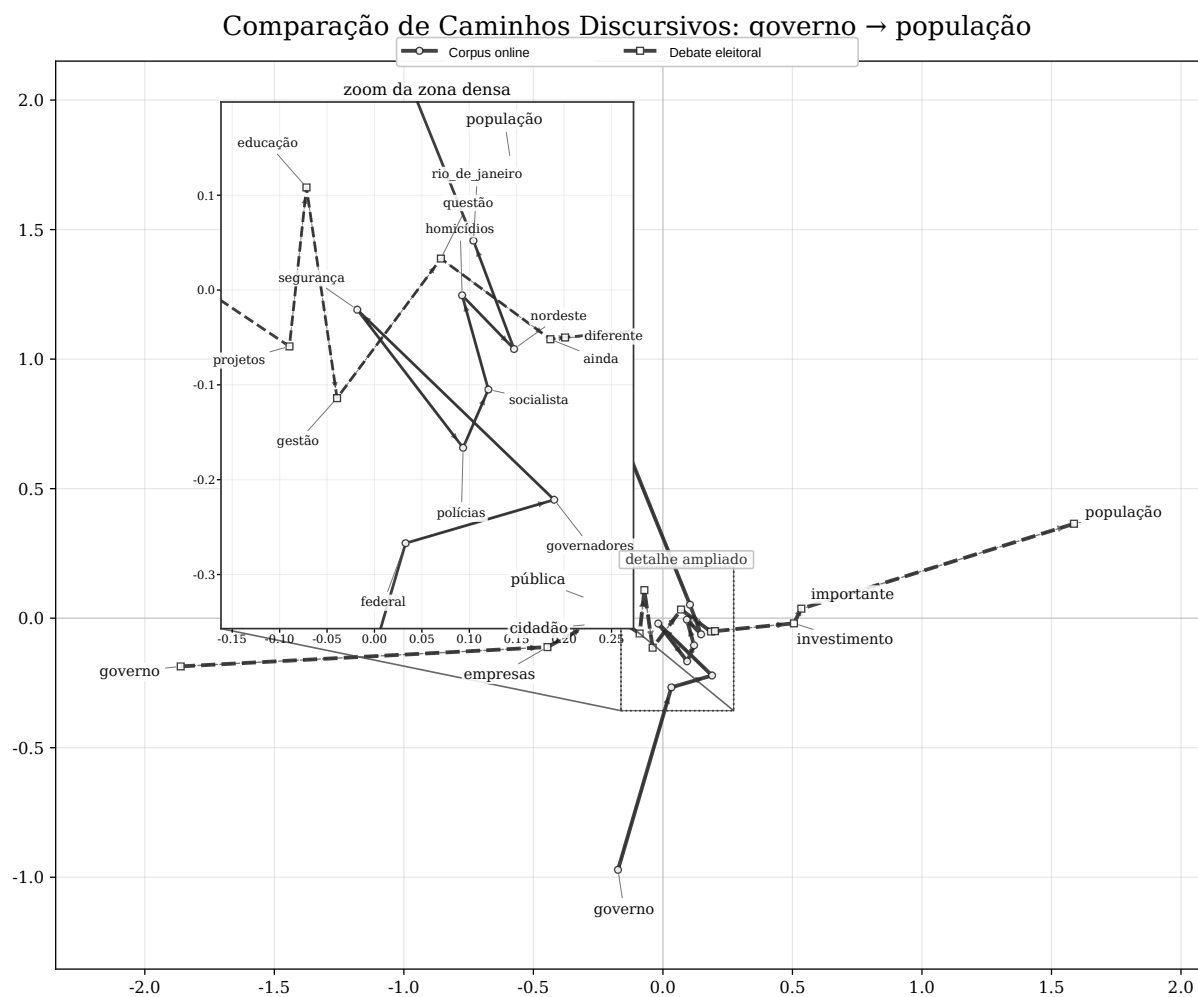


Figura 3: “Governo” → “População”, gráfico comparativo entre corpora, com detalhe ampliado da zona densa.

ALT: Gráfico de dispersão comparando os caminhos discursivos para o par governo–população em dois corpora. O corpus do debate apresenta intermediários ligados a empresas, cidadania, política pública, projetos, educação, gestão e investimento; o corpus online apresenta intermediários ligados a federalismo, governadores, segurança, polícias, homicídios, Nordeste e Rio de Janeiro.

Debate governo → empresas → cidadão → pública → projetos → educação → gestão →
questão → ainda → diferente → investimento → importante → população

Online governo → federal → governadores → segurança → polícias → socialista → homi-
cídios → nordeste → rio_de_janeiro → população

No corpus do debate, “governo” e “população” são ligados por um inventário de política pública: “empresas”, “cidadão”, “pública”, “projetos”, “educação”, “gestão”, “investimento” e “importante”. A população aparece no fim como destinatária de ações, projetos e prioridades. A presença de “empresas” logo depois de “governo” é reveladora, porque introduz uma mediação público-privada antes mesmo que surja “cidadão”. Em seguida, “educação”, “gestão” e “investimento” organizam o caminho como vocabulário de administração, desenho de política e entrega institucional. Itens como “ainda” e “diferente” são pouco interpretáveis isoladamente, mas funcionam como partículas de transição em uma cadeia cujo núcleo semântico é suficientemente claro: o governo chega à população pela linguagem da gestão.

No corpus online, o mesmo par é conectado por um repertório mais áspero: “federal”, “governadores”, “segurança”, “polícias”, “socialista”, “homicídios”, “nordeste” e “rio_de_janeiro”. O poder público aparece distribuído em níveis federativos e atores concretos; em seguida, a passagem por “segurança”, “polícias” e “homicídios” desloca a população do lugar de beneficiária da gestão para o de corpo exposto à coerção, ao medo e à morte. “nordeste” e “rio_de_janeiro” territorializam o trajeto: a população não é apenas uma categoria estatística, mas um lugar social situado, atravessado por regionalização, policiamento e disputa ideológica. O contraste local, portanto, é entre governo como administração de temas e governo como aparato cujos efeitos se materializam em violência, território e conflito. 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”

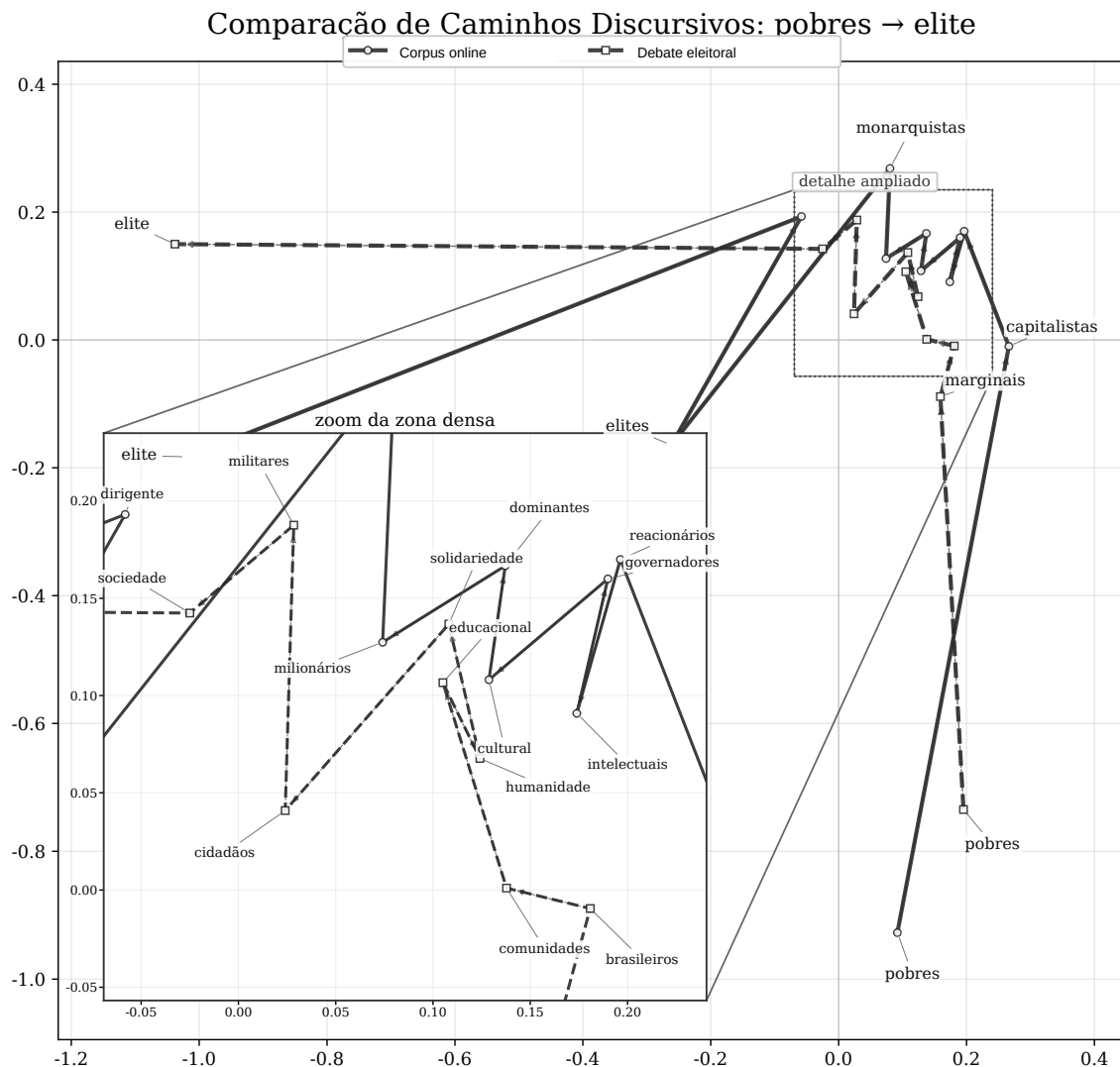


Figura 4: “Pobres” → “Elite”, gráfico comparativo entre corpora, com detalhe ampliado da zona densa.

ALT: Gráfico de dispersão comparando os caminhos discursivos para o par pobres–elite em dois corpora. O corpus do debate apresenta intermediários ligados a marginalização, brasileiros, comunidades, educação, humanidade, solidariedade, cidadania, militares e sociedade; o corpus online apresenta intermediários ligados a capitalistas, reacionários, intelectuais, governadores, dominantes, milionários, monarquistas e elites dirigentes.

Debate	pobres → marginais → brasileiros → comunidades → educacional → humanidade → solidariedade → cidadãos → militares → sociedade → elite
---------------	---

Online	pobres → capitalistas → reacionários → intelectuais → governadores → cultural → dominantes → milionários → monarquistas → elites → dirigente → elite
---------------	---

Este é o caso em que a assimetria entre os corpora se torna mais sensível. No debate, a transição de “pobres” para “elite” começa com “marginais”, o que introduz um estigma inicial. Contudo, o percurso não permanece nesse registro. Ele se desloca para “brasileiros”, “comunidades”, “educacional”, “humanidade”, “solidariedade”, “cidadãos” e “sociedade”. A cadeia oscila, portanto, entre marginalização e integração. Primeiro, os pobres são aproximados do desvio; depois, são reinseridos por uma gramática comunitária, pedagógica e humanitária. A presença de “militares” perto de “sociedade” acrescenta uma camada de ordem institucional: a passagem rumo à elite não é construída como antagonismo de classe, mas como movimento através de pertencimento nacional, educação, solidariedade, cidadania e disciplina social.

No corpus online, a lógica é outra. “pobres” vai imediatamente a “capitalistas”, “reacionários”, “intelectuais”, “governadores”, “dominantes”, “milionários”, “monarquistas”, “elites”, “dirigente” e “elite”. A distância social é preenchida por nomes de posição, classe e comando. “cultural” é particularmente interessante porque impede uma leitura puramente econômica: entre “governadores” e “dominantes”, ele sugere que a dominação aparece também como disputa simbólica, não apenas patrimonial. Há aqui uma “representação antagonística de grupos sociais” (PIVETTA; GONÇALVES-SEGUNDO, 2024, p. 18) pela nomeação progressivamente mais explícita de adversários, frações dirigentes e imaginários de poder. O contraste local mostra um debate que integra a pobreza por cidadania, comunidade e ordem, contra um corpus online que reconstrói a passagem até a elite como mapa de dominação. 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”

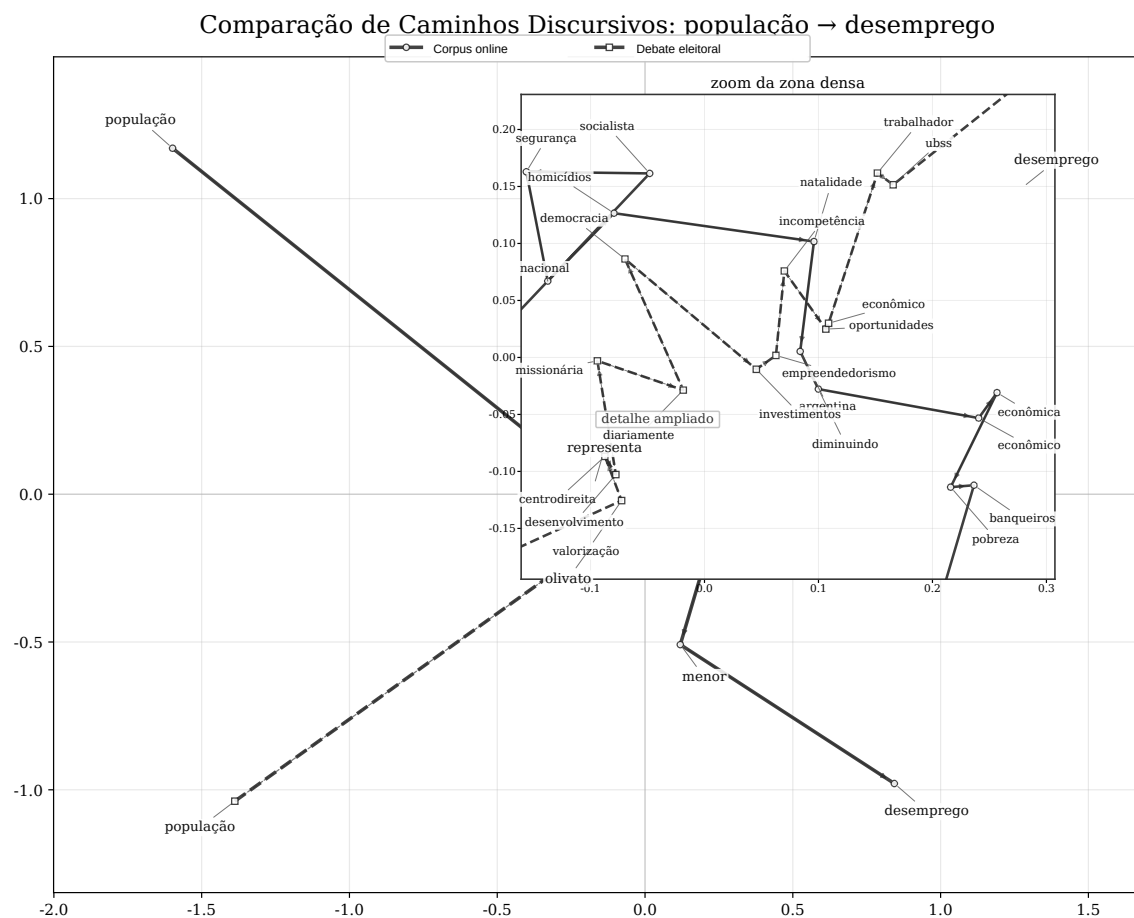


Figura 5: “População” → “Desemprego”, gráfico comparativo entre corpora, com detalhe ampliado da zona densa.

ALT: Gráfico de dispersão comparando os caminhos discursivos para o par população–desemprego em dois corpora. O corpus do debate apresenta intermediários ligados a valorização, centro-direita, desenvolvimento, democracia, investimentos, empreendedorismo, incompetência, oportunidades, trabalhador e serviço público; o corpus online apresenta intermediários ligados a socialismo, segurança nacional, homicídios, natalidade, Argentina, economia, pobreza, banqueiros e menor desemprego.

Debate	população → olivato → valorização → centrodireita → desenvolvimento → missionária → diariamente → democracia → investimentos → empreendedorismo → incompetência → oportunidades → econômico → trabalhador → ubss → desemprego
---------------	---

Online	população → representa → socialista → segurança → nacional → homicídios → natalidade → argentina → diminuindo → econômico → econômica → pobreza → banqueiros → 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 “valorização”, “centrodireita”, “desenvolvimento”, “democracia”, “investimentos”, “empreendedorismo”, “incompetência”, “oportunidades”, “econômico”, “trabalhador” e “ubss”. A cadeia sugere uma causalidade gerencial e moralizada: o desemprego é concebido como problema de desenvolvimento, investimento, empreendedorismo, oportunidade e trabalhador, mas também como efeito de condução ruim ou “incompetência”. “centrodireita” introduz um marcador ideológico explícito, enquanto “ubss” recoloca, no fim do caminho, o desemprego dentro de uma paisagem de serviços públicos. Há também itens idiossincráticos, como “olivato”, “missionária” e “diariamente”; eles recomendam cautela, mas não anulam a inteligibilidade do núcleo semântico.

No corpus online, a rota é mais estrutural e mais inquieta. “população” passa por “socialista”, “segurança”, “nacional”, “homicídios”, “natalidade”, “argentina”, “econômico”, “econômica”, “pobreza”, “banqueiros” e “menor” antes de chegar a “desemprego”. O desemprego não aparece como falha administrativa pontual, mas como efeito distribuído entre segurança pública, demografia, comparação econômica regional, pobreza e sistema financeiro. “argentina” funciona como índice comparativo e imaginário de crise; “banqueiros” recoloca a cadeia no terreno dos agentes financeiros; “pobreza” faz a ponte entre economia e experiência social. Esse caso importa para o padrão geral porque condensa a oposição central da seção: no debate, o problema social é filtrado por desenvolvimento, empreendedorismo, investimento e empregabilidade; no corpus online, ele é reancorado em insegurança, pobreza, demografia e agentes financeiros reconhecíveis.

5.6 Recorrência, alcance e limites

O contraste não é uniforme. Em *pobres–elite*, “marginais” introduz no debate uma associação estigmatizante, e “militares” tensiona a sequência humanitária: há conflito explícito,

não apenas amortecimento. Em *democracia–ditadura*, “comunista” aparece nos dois corpora, e “república” também integra ambas as rotas. No corpus online, “federal” e “governadores” mantêm um repertório institucional, enquanto “benefício” e “famílias”, em *dinheiro–elite*, não implicam por si mesmos acusação ou antagonismo. Esses casos impedem opor dois regimes puros e recomendam tratar as categorias como tendências locais. A presença de vocabulário administrativo pode decorrer do tema discutido, assim como a proximidade de palavras de conflito pode decorrer de crítica, citação ou negação; a geometria, isoladamente, não distingue essas funções.

Tomados em conjunto, os cinco pares focalizados permitem uma síntese contrastiva restrita aos casos e às configurações examinados. No corpus do debate, repetem-se intermediários de gestão, serviços públicos, regulação, institucionalidade, cidadania administrativa e solução técnico-política. No corpus online de referência, repetem-se intermediários ligados a crise democrática, acusação política, antagonistas de classe, economia cotidiana, violência, território, pobreza e atores financeiros; esse padrão ajuda a tratar os caminhos discursivos geométricos como “formas produtivas de nomeação e predicação” (PIVETTA; GONÇALVES-SEGUNDO, 2024, 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, 2024, 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, 2024, p. 13) e podem levar à “estagnação e não desenvolvimento de argumentos” (TEIXEIRA; DAMASCENO-MORAIS, 2024, 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 judiciário 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 argumentativamente responsável do ponto central do artigo: nos cinco pares focalizados, observa-se recorrência de intermediários técnico-administrativos, regulatórios, institucionalizantes e moral-gerenciais no corpus do debate, em contraste com intermediários mais historicamente ancorados, socialmente concretos, antagonísticos, securitários ou econômico-financeiros no corpus online de referência. Os critérios de verificação complementar de estabilidade, sensibilidade ao tamanho do corpus e comparação com versões menos restritivas do procedimento são explicitados no material suplementar. No corpo do artigo, o ponto decisivo permanece 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 O que os caminhos mostram sobre o que o discurso esconde

O centro do argumento está aqui: os caminhos discursivos geométricos oferecem indícios para examinar como um discurso pode ocultar não por apagar uma palavra, mas por interpor uma cadeia. Essa cadeia não é detalhe técnico. Ela pode participar da forma como o discurso faz uma relação social parecer natural, distante, administrável ou inevitável. Por isso, os caminhos continuam sendo decisivos: não são uma ilustração lateral, mas a evidência que permite passar de uma intuição crítica para uma análise discutível.

Nos cinco pares focalizados, o corpus do debate tende a construir passagens por mediações institucionalmente legítimas. Democracia e ditadura passam por regulação, comunicação, criminalidade e classificação ideológica; dinheiro e elite passam por política, compromisso, economia e empresarização; governo e população passam por empresas, projetos, educação, gestão e investimento; população e desemprego passam por desenvolvimento, empreendedorismo, oportunidades e serviço público. O padrão não é que o debate seja “falso”. O padrão é que ele frequentemente torna o conflito menos direto. A política aparece como uma cadeia de gestão. A desigualdade aparece como economia. O desemprego aparece como empregabilidade. A crise democrática aparece como disputa de comunicação e mentalidade.

Essa observação exige equilíbrio. O vocabulário técnico não é ruim por definição. Políticas públicas precisam de palavras como gestão, investimento, regulação, educação, serviço

e planejamento. O problema começa quando essas palavras passam a ocupar sistematicamente a posição de ponte entre conceitos que carregam conflito social. Nesse caso, a mediação técnica não apenas esclarece; ela também protege. Ela torna certas relações menos nomeáveis. Chamo esse efeito, de modo provisório, de *autoproteção semântica*: uma forma de o discurso se defender de suas próprias implicações por meio de intermediários legítimos.

No corpus online de referência, o padrão é outro. As rotas tendem a acionar impeachment, golpe, massacre, banqueiros, milionários, capitalistas, governadores, polícias, homicídios, pobreza e território. A política aparece menos filtrada pela linguagem da gestão e mais atravessada por eventos, atores e antagonismos nomeáveis. Isso não torna o corpus online mais verdadeiro, mais puro ou mais democrático. Ao contrário, a circulação digital também pode simplificar, acusar, agredir e produzir cadeias de antagonização que empobrecem a argumentação (OLIVEIRA; MIRANDA, 2024; OLIVEIRA et al., 2025; PIVETTA; GONÇALVES-SEGUNDO, 2024; TEIXEIRA; DAMASCENO-MORAIS, 2024). O contraste, portanto, não é entre instituição enganosa e internet autêntica. É entre dois modos de tornar o conflito linguisticamente disponível.

Essa distinção importa especialmente em 2026. A agenda pública sobre eleições, inteligência artificial e desinformação tende a privilegiar conteúdos fabricados, conteúdos manipulados, deepfakes e rotulagem de material sintético, tema que já aparece na comunicação institucional do TSE para as Eleições Gerais de 2026 (TRIBUNAL SUPERIOR ELEITORAL, 2026a; TRIBUNAL SUPERIOR ELEITORAL, 2026b). Essa preocupação é necessária. Mas ela não esgota o problema linguístico. A Linguística precisa insistir em uma camada menos visível: mesmo quando não há falsificação factual evidente, o discurso pode reorganizar a relação entre conceitos de modo a deslocar responsabilidades, amortecer antagonismos ou transformar conflitos sociais em problemas de administração.

É por isso que um corpus de 2024 pode falar ao presente. O material vem de uma eleição municipal anterior, mas a pergunta que ele abre é atual: se as eleições de 2026 serão disputadas em torno de segurança, economia, desigualdade, democracia, instituições, plataformas e desinformação, então não basta observar temas e checar fatos. É preciso observar os caminhos. Que pontes vão ligar *democracia a ordem*? *Pobreza a mérito*? *Desemprego a empreendedorismo*? *Elite a eficiência*? *População a segurança*? A disputa política acontece também nesses elos.

A contribuição para a Linguística textual está em deslocar a unidade de observação. A palavra isolada continua importante, mas não basta. A frequência continua útil, mas não basta. A classificação de polaridade ou de tópico continua relevante, mas não basta. O que os caminhos acrescentam é uma camada mesoestrutural: entre a palavra e o discurso inteiro, há trajetos de

mediação. Eles permitem investigar quais intermediários tornam plausível a conexão entre polos, sem presumir que a sequência calculada tenha sido enunciada ou percorrida pelo leitor.

Para a análise do discurso, o ganho é outro. A técnica ajuda a formular melhor a crítica. Em vez de dizer apenas que um discurso é tecnocrático, podemos perguntar onde a tecnocratização aparece no trajeto semântico. Em vez de dizer apenas que o discurso online é agressivo, podemos perguntar quais antagonistas, eventos e agentes materiais estruturam as passagens entre conceitos. Em vez de opor qualitativo e computacional, podemos tratar o cálculo como instrumento de localização de problemas interpretativos. O algoritmo não interpreta no lugar do analista; ele aponta regiões em que a interpretação precisa trabalhar.

Para a pesquisa futura, há pelo menos quatro caminhos. Primeiro, ampliar o método para outros gêneros: debates nacionais, programas de governo, propaganda eleitoral, entrevistas, posts de campanha e material sintético produzido por inteligência artificial. Segundo, comparar subcorpora por candidato, partido, plataforma e momento da campanha, para verificar se a autoproteção semântica aparece como efeito de gênero ou como estratégia localizada. Terceiro, articular os caminhos discursivos com análise qualitativa de sequências enunciativas, evitando que a visualização substitua a leitura. Quarto, incorporar esses procedimentos a práticas de transparência e ciência aberta, permitindo que outros pesquisadores refaçam, contestem ou ampliem as rotas.

O ponto mais importante, porém, é teórico. Caminhos discursivos geométricos não servem apenas para dizer que dois corpora são diferentes. Servem para perguntar o que essa diferença faz. Se um caminho entre *dinheiro* e *elite* passa por *economia* e *empresarização*, a relação de classe é reconfigurada de um modo. Se passa por *patrimônio*, *banqueiros* e *milionários*, é reconfigurada de outro. O método interessa porque torna essas pontes visíveis. A crítica começa quando perguntamos por que elas estão ali.

Assim, a questão maior não é metodológica no sentido estreito. É política e linguística. Em ano eleitoral, o que ameaça a qualidade do debate público não é apenas a mentira explícita, mas também a naturalização silenciosa de trajetos semânticos que fazem certos conflitos desaparecerem dentro de palavras legítimas. A Linguística tem instrumentos para examinar essa camada. Os caminhos discursivos geométricos são uma proposta de fazê-lo de forma controlada, auditável e aberta à discordância.

7 Conclusão

Este artigo defendeu que uma parte decisiva do funcionamento político do discurso está nos caminhos que ligam conceitos, e não apenas nos conceitos tomados isoladamente. A noção de caminhos discursivos geométricos foi proposta para tornar essa camada mais visível. Seu valor não está em substituir a interpretação textual por cálculo, mas em localizar pontes semânticas que podem ser discutidas teoricamente.

A leitura dos cinco pares focalizados sustenta, no escopo do caso analisado, uma diferença relevante entre dois campos simbólicos. No debate eleitoral, conceitos politicamente densos tendem a ser ligados por mediações institucionalmente legítimas, técnico-administrativas ou moral-gerenciais. No corpus de comentários online de referência, os mesmos pares tendem a acionar eventos, antagonistas nomeados, agentes financeiros, memória histórica e efeitos sociais materializados. Essa diferença não deve ser lida como prova de que um corpus seria mais verdadeiro do que o outro. Ela sustenta a hipótese de que, nesses casos, repertórios distintos podem tornar conflitos mais ou menos nomeáveis por trajetos semânticos diferentes. Como escala, composição e configuração dos modelos variam simultaneamente, o contraste não isola efeitos de gênero nem constitui prova automática de ocultação.

A contribuição disciplinar está nessa mudança de foco. Em vez de apresentar a técnica como solução externa para problemas de interpretação, o artigo propõe que a técnica seja avaliada por sua capacidade de tornar melhores as perguntas teóricas: que mediações um texto normaliza? Que conflitos ele desloca para a linguagem da gestão? Que relações sociais ele faz aparecer como administração, oportunidade, segurança ou eficiência? Que responsabilidade interpretativa acompanha a passagem de uma visualização vetorial para uma afirmação sobre discurso?

Essa agenda é especialmente relevante no ciclo eleitoral de 2026. A discussão pública sobre inteligência artificial, deepfakes e desinformação é indispensável, mas não basta para a Linguística. Também é preciso examinar discursos que não são necessariamente falsos, mas que organizam caminhos capazes de amortecer conflitos, deslocar responsabilidades ou proteger certas relações de força. O que se oculta no discurso pode estar justamente no meio do caminho.

Como agenda de pesquisa, interessa ampliar o teste para corpora maiores, outros gêneros institucionais e desenhos comparativos mais balanceados, com arquitetura, janela e critérios lexicais comuns, além de refinar limiares de persistência e procedimentos de robustez. Também será necessário examinar pares escolhidos por critérios independentes, comparar subcorpora por falante ou por período e discutir como esse tipo de métrica pode dialogar com

análises qualitativas de argumentação, impolidez, desinformação, tecnocratização e memória discursiva. O caminho proposto, portanto, não encerra a questão; oferece uma forma documentada de recolocá-la.

Conflito de interesses

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

Nota ética e normativa

O estudo utiliza materiais textuais de acesso público: transcrições de debates eleitorais transmitidos publicamente e comentários online coletados sem 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 estarem sujeitos a restrições autorais e de licenciamento das emissoras e das plataformas de hospedagem.

Declaração de disponibilidade de dados

Os dados, materiais e rotinas analíticas que sustentam a evidência mobilizada neste 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. Estão incluídos o script principal de análise, as rotinas auxiliares de coleta, limpeza e avaliação, as matrizes ajustadas aos dois corpora e o manifesto dos caminhos. O recurso fastText de origem é obtido por download automático do endereço oficial indicado no script. As dependências, a preparação do ambiente e a distinção entre reutilização dos caminhos preservados e recálculo são descritas no Apêndice, em “Condições de execução e reprodução”. Endereço do repositório: <https://github.com/alexandre-barroso/GeometriaDiscursiva>. O identificador persistente (DOI) deste repositório é <https://doi.org/10.5281/zenodo.19392318> (licença de uso: CC BY 4.0). Conforme informado anteriormente, 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 estarem sujeitos a restrições autorais e de licenciamento das emissoras e das plataformas de hospedagem. Todos os debates estão

disponíveis abertamente pelo YouTube, nos canais oficiais das emissoras e/ou suas plataformas.

Fonte de financiamento

O autor declara ter recebido bolsa doutoral vigente durante a realização desta pesquisa. CAPES, número de processo: 88887.186643/2025-00.

Declaração de uso de IA

A Inteligência Artificial de transcrição automática (áudio para transcrição textual) *Whisper* da empresa OpenAI foi utilizada, em um primeiro momento para transformar os áudios dos debates em transcrições textuais com contribuição mínima. Todo o conteúdo foi revisado e editado pelo pesquisador, que assume total responsabilidade pelo manuscrito, isto é, todas as transcrições geradas por IA foram manualmente checadas e organizadas pelo pesquisador, como relatado na seção metodológica do artigo.

Contribuição de Autoria

Alexandre Menezes Barroso: Conceitualização, Investigação, Metodologia, Software, Análise Formal, Visualização, Escrita – esboço original, Escrita – revisão e edição.

Link para *preprint*

Link para Preprint <https://doi.org/10.1590/SciELOPreprints.15726>

Referências bibliográficas

BARROSO, Alexandre. Geometrias discursivas V1.0.0: versão da submissão. Zenodo, 2026. DOI: <https://doi.org/10.5281/zenodo.19392318>. Acesso em: 5 jun. 2026.

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: Editora 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.

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.2025.100929.

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 metadiscorso 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.

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. Espriamento 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.

TRIBUNAL SUPERIOR ELEITORAL. **Eleições**. Brasília, DF: TSE, 2026a. Disponível em: <https://www.tse.jus.br/eleicoes>. Acesso em: 5 jun. 2026.

TRIBUNAL SUPERIOR ELEITORAL. **Por Dentro das Eleições: conheça as regras sobre uso de IA na campanha eleitoral de 2026**. Brasília, DF: TSE, 10 abr. 2026b. Disponível em: <https://www.tse.jus.br/comunicacao/noticias/2026/Abril/por-dentro-das-eleicoes-conheca-as-regras-sobre-uso-de-ia-na-campanha-eleitoral-de-2026>. Acesso em: 5 jun. 2026.

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 nem deslocar o centro da submissão para o reporte de resultados, mas tornar auditável a evidência que sustenta o argumento principal. Em coerência com o contrato do manuscrito, mantemos em primeiro plano a pergunta de Linguística textual e discursiva: que mediações tornam certas relações políticas mais visíveis, mais indiretas ou menos nomeáveis?

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 com dados do *Common Crawl* e da Wikipédia, 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 de ajuste (10) e o procedimento de ajuste sobre o vocabulário efetivamente presente em cada corpus. A rotina realiza cinco épocas preliminares de Word2Vec no corpus, substitui os vetores de entrada dos itens compartilhados pelos vetores *fastText* e executa dez épocas adicionais de ajuste, com todos

os vetores liberados para atualização. Os itens sem correspondência no recurso inicial mantêm a representação obtida na etapa preliminar. Trata-se de transferência de vetores lexicais, não de continuação do treinamento sublexical do fastText; a procedência e os parâmetros do recurso inicial são descritos na Seção 3.

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. CBOW prevê uma palavra a partir do contexto; Skip-gram prevê contextos a partir de uma palavra. Essa diferença pode afetar as representações, inclusive de itens raros, e interfere tanto na seleção quanto na ordenação dos intermediários. A comparação de rotas continua descritivamente útil, mas não controla o efeito da arquitetura, da janela ou da frequência mínima. Assim, as diferenças não são atribuídas exclusivamente aos corpora ou ao gênero discursivo.

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 distribucional.

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, removeram-se as *stopwords* da lista de português do NLTK, com as exceções indicadas a seguir. Segundo, preservaram-se deliberadamente algumas palavras funcionais com potencial valor discursivo, como *não*, *sim*, *muito*, *pouco*, *mais* e *menos*. Terceiro, exigiu-se a presença de caracteres compatíveis com o padrão alfabético definido na rotina. Quarto, descartaram-se tokens de um caractere, salvo *é*, *à*, *e*, *o* e *a*, desde que tivessem passado pelos filtros anteriores; essa lista não reinsere palavras já removidas como stopwords. Não houve lematização, etiquetagem morfosintática nem reconhecimento automático de entidades nomeadas. O procedimento é lexical-distribucional e opera sobre coocorrência. O recurso pré-treinado inclui palavras funcionais; a filtragem por nossa lista de stopwords pertence à adaptação aos corpora. Ao excluir esses tokens antes de formar as janelas, aproximamos palavras de conteúdo antes mais distantes e alteramos seus contextos de treinamento. Essa decisão não é neutra nem é anulada pela

inicialização comum: as rotas descrevem os corpora após essa filtragem, e não a ordem lexical integral dos enunciados.

As linhas vazias foram descartadas, e apenas unidades com mais de três tokens após o pré-processamento foram efetivamente aproveitadas no treinamento. As unidades são as linhas dos arquivos consolidados: uma no corpus dos debates e sete no corpus online. As janelas são locais dentro dessas linhas, mas não preservam necessariamente fronteiras de turnos, comentários ou períodos sintáticos. As matrizes ajustadas contêm, respectivamente, 4.728 e 22.833 itens; a diferença de cobertura se soma à assimetria de volume indicada na Seção 4.

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 SVD completa (*svd_solver=full*), gerando uma base ortogonal completa para projeções sucessivas do campo semântico.

O cálculo da persistência percorre 299 projeções dimensionais decrescentes, de 300 até 2 dimensões. Em cada projeção, calcula-se o ponto médio entre os âncoras e selecionam-se os dez itens mais próximos por distância euclidiana, excluindo os próprios âncoras da candidatura. Sejam a_d e b_d os âncoras e x_d o candidato na projeção de dimensão d . Definem-se

$$L_d = \|b_d - a_d\|_2, \quad s_d(x) = \left\langle x_d - a_d, \frac{b_d - a_d}{L_d} \right\rangle.$$

Na rotina, *path_len* corresponde a L_d , o comprimento do segmento, e *proj_len* a $s_d(x)$, a projeção escalar orientada desde o primeiro âncora; *len* abrevia *length* (comprimento). Retém-se o candidato quando $0 < s_d(x) < L_d$. Não há um comprimento fixo adotado: ambos os valores são calculados para cada par e projeção. O teste delimita a posição longitudinal, não exige que o vetor esteja sobre o segmento nem fixa uma distância lateral máxima. Projeções com $L_d < 10^{-12}$ são descartadas para evitar divisão por comprimento numericamente nulo.

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 os itens com contagem superior a $0,3 \times 300 = 90$, isto é, ao menos 91 ocorrências nas 299 projeções consideradas. A base operacional do limiar é a dimensionalidade inicial, 300. O critério de persistência, e não uma garantia estatística associada a esse limiar heurístico, constitui 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, calcula-se $\bar{s}(x) = \frac{1}{5} \sum_{d \in \{300, 250, 200, 150, 100\}} s_d(x)$ e ordenam-se os candidatos por valores crescentes. Não se calcula a distância ao ponto médio nessa etapa, nem se divide cada projeção por L_d . Eventuais projeções descartadas por comprimento numericamente nulo não entram na média. 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. Os empates de distância são resolvidos por ordem lexical. Na ordenação final, as posições médias são arredondadas a 12 casas decimais e, em caso de empate, aplica-se novamente a ordem lexical. A repetibilidade numérica de uma execução não deve ser confundida com a validade interpretativa 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 <i>fastText</i> para o português do <i>Common Crawl</i> e da Wikipédia.
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 de ajuste	10, após cinco épocas preliminares.
Amostragem negativa	15.
Taxa de aprendizado	Inicial 0.025; mínima 0.0001; atualização a cada época conforme a rotina.
Tokenização	Tokenização em português; caixa baixa; filtragem lexical.
Stopwords	Removidas com preservação deliberada de <i>não</i> , <i>sim</i> , <i>muito</i> , <i>pouco</i> , <i>mais</i> e <i>menos</i> .
Filtros adicionais	Filtro de caracteres e comprimento; sem reinclusão de stopwords; descarte de linhas vazias e de unidades com menos de quatro tokens.
Base de projeção	PCA por SVD completa sobre o espaço centralizado.
Semente fixa	42 no treinamento; um trabalhador e uma thread numérica.
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	Projeção escalar $0 < s_d(x) < L_d$; exclusão dos âncoras.
Regra de persistência	Ao menos 91 ocorrências nas 299 projeções.
Ordenação final	Projeção escalar média desde o primeiro âncora 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

Condições de execução e reprodução

O repositório indicado na declaração de disponibilidade reúne os corpora, as rotinas auxiliares, o script principal, as matrizes ajustadas e o manifesto dos caminhos. As matrizes preservadas registram Python 3.12.3 e Gensim 4.3.3. A execução utiliza também NumPy, SciPy, NLTK, scikit-learn e Matplotlib, em CPU, com semente 42, um trabalhador e uma thread por biblioteca numérica. Na pasta do script, mantêm-se os diretórios `corpora` e `models`; cria-se previamente `pretrained_models`. O NLTK requer os recursos de português `stopwords` e `punkt`, além de `punkt_tab` nas versões que o utilizam.

O script obtém automaticamente `cc.pt.300.vec.gz` do endereço oficial nele registrado e o descomprime. Esse recurso de origem, distinto das matrizes ajustadas aos corpora, requer conexão para o download, espaço em disco para as versões compactada e descompactada e memória para carregar os vetores e as matrizes de projeção. O script verifica sua presença antes de carregar os modelos locais. Por padrão, reutiliza as matrizes e o manifesto; para recalcular caminhos, define-se `GEOMETRIAS_REBUILD_PATHS=1`; para retreinar os modelos e recalcular as rotas, define-se também `GEOMETRIAS_FORCE_RETRAIN=1`. Reutilizar o manifesto reproduz as sequências preservadas, mas não constitui teste independente de estabilidade. A identidade numérica entre execuções exige os mesmos dados, recursos de tokenização, versões das bibliotecas e ambiente numérico. A reprodução da análise parte dos corpora preservados; uma nova coleta na plataforma constitui outra amostra.

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, pre-

servando 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. Um exemplo 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 “corpus online de referência”, 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, executada por comunidade, consulta até 1.000 tópicos na ordenação *hot*, percorre os selecionados em ordem aleatória e registra títulos e quantidades de comentários por fio. A chamada `replace_more`, com `limit=0`, não expande blocos adicionais: são armazenados os corpos dos comentários retornados, sem garantia de recuperar a totalidade de cada discussão. Os arquivos consolidados não mantêm a hierarquia das respostas. Para repetir a coleta, a rotina auxiliar utiliza PRAW e credenciais próprias da API; o treinamento sobre os corpora preservados dispensa esse acesso.

O material foi então consolidado em um único arquivo e submetido a limpeza mo-

derada: 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	<i>rbrasil, rbrasilivre, rinvestimentos, reddit</i>
Risada gráfica e ruído interacional	reduzir repetição gráfica sem conteúdo temático robusto	<i>kkk, kkkkk, rs, rsrs, haha, hahaha</i>
Abreviações conversacionais muito circunstanciais	diminuir dispersão causada por marcas interacionais pouco estáveis	<i>pfv, pls, vdd, oq, sqn, flw, vlw</i>
Regularização ortográfica e acentual	aproximar variantes evidentes de uma mesma unidade lexical	<i>politica → política; educa-cao → educação; familia → família</i>
Expansão de abreviações frequentes	concentrar coocorrências hoje dispersas em formas abreviadas	<i>vc → você; cmg → comigo; pra → para</i>
Unificação temática e multiword	preservar entidades e tópicos recorrentes como unidades semânticas estáveis	<i>rio de janeiro → rio_de_janeiro; são paulo → são_paulo; covid/corona → coronavírus</i>

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 maior escala dos comentários amplia a cobertura lexical e pode estabilizar representações pouco estimadas no debate; a composição temática altera quais palavras competem pelo ponto médio. Em contrapartida, agregar fóruns ou falantes reúne usos divergentes em um mesmo vetor. A seleção pela plataforma, a limpeza diferenciada e as janelas aplicadas após remoção de stopwords também afetam as rotas. Mesmo a redução de volume, isoladamente, não controlaria essas diferenças nem as arquiteturas de treinamento.

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 distribucional. 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 (<i>bom–ruim, feliz–triste, rico–pobre</i> 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 delimita três frentes para ampliar a validação: estabilidade das rotas em reexecuções, sensibilidade ao tamanho do corpus online e comparação com versões menos restritivas do procedimento. Neste suplemento, essas frentes são apresentadas como critérios de verificação, não como resultados experimentais. A persistência dimensional verifica recorrência entre projeções de um mesmo modelo e não substitui reamostragem, variação de sementes ou comparação sob parâmetros comuns. 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 as famílias de intermediários observadas permanecem sob perturbações do procedimento, registrando também mudanças, aproximações entre corpora e inversões do contraste, em vez de exigir confirmação da hipótese inicial.

Estabilidade entre reexecuções

Em procedimentos de treinamento distribucional, alterações de semente ou ambiente numérico podem produzir variações locais. A rotina empregada fixa a semente global em 42 e controla o paralelismo, o que favorece repetibilidade sob as mesmas condições, mas não testa estabilidade perante outras configurações. 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. Essa expectativa precisa ser examinada empiricamente: mais candidatos não significam, por si só, menor pertinência discursiva.

O limiar descrito na especificação foi adotado como compromisso heurístico 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 a composição e a interpretação da rota resistem a verificações independentes, inclusive quando revelam convergências entre corpora ou contradizem a oposição inicialmente proposta. 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 Apoio documental dos cinco caminhos

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 → república → ubss → arsesp → obviamente → roman- tizando → comunicação → crimi- nalidade → mentalidade → comu- nista → esquerdista → ditadura	democracia → impeachment → di- tatorial → liberais → democrática → golpista → massacre → repú- blica → comunista → ditadura
dinheiro–elite	dinheiro → população → real- mente → rio_de_janeiro → histó- ria → política → compromisso → principal → fundamental → econo- mia → empresarização → princi- palmente → elite	dinheiro → empréstimo → bilio- nário → tesouro → financiamento → benefício → patrimônio → fa- mílias → milionário → exclusiva- mente → banqueiros → milioná- rios → elite

Par	Corpus do debate	Corpus online de referência
governo– população	governo → empresas → cidadão → pública → projetos → educa- ção → gestão → questão → ainda → diferente → investimento → im- portante → população	governo → federal → governado- res → segurança → polícias → so- cialista → homicídios → nordeste → rio_de_janeiro → população
pobres–elite	pobres → marginais → brasileiros → comunidades → educacional → humanidade → solidariedade → ci- dadãos → militares → sociedade → elite	pobres → capitalistas → reacioná- rios → intelectuais → governado- res → cultural → dominantes → milionários → monarquistas → eli- tes → dirigente → elite
população– desemprego	população → olivato → valoriza- ção → centrodireita → desenvol- vimento → missionária → diari- amente → democracia → investi- mentos → empreendedorismo → incompetência → oportunidades → econômico → trabalhador → ubss → desemprego	população → representa → socia- lista → segurança → nacional → homicídios → natalidade → argen- tina → diminuindo → econômico → econômica → pobreza → ban- queiros → 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	república; ubss; arsesp; comunicação; criminalidade; comunista; esquerdista	mediação institucional, comunicacional e classificação ideológica	conflito reencaminhado por serviços, regulação e avaliação do adversário
democracia–ditadura	Online	impeachment; ditatorial; golpista; massacre; república; comunista	ruptura política e memória de regime	a passagem reúne eventos, acusações e formas políticas
dinheiro–elite	Debate	população; política; compromisso; economia; empresarização	legitimação institucional e enquadramento econômico	a elite aparece mediada por linguagem pública e organização econômica
dinheiro–elite	Online	empréstimo; tesouro; financiamento; patrimônio; famílias; banqueiros; milionários	crédito, acumulação e agentes financeiros	o caminho articula operações financeiras e concentração de riqueza
governo–população	Debate	empresas; cidadão; pública; projetos; educação; gestão; investimento	inventário de gestão e política pública	a população emerge como destinatária de ações administráveis
governo–população	Online	federal; governadores; segurança; polícias; homicídios; nordeste; rio_de_janeiro	poder estatal, territorialização e violência	a mediação passa por aparato público e efeitos sociais
pobres–elite	Debate	marginais; comunidades; educacional; humanidade; solidariedade; cidadãos; militares; sociedade	tensão entre estigma, integração e ordem	o trajeto associa marginalização e pertencimento social
pobres–elite	Online	capitalistas; reacionários; intelectuais; governadores; cultural; dominantes; milionários; elites	antagonismo de classe e dominação simbólica	o conflito é interpretado pela nomeação de posições sociais

Par	Corpus	Intermediários de maior peso interpretativo	Categoria funcional predominante	Observação
população–desemprego	Debate	desenvolvimento; investimentos; empreendedorismo; incompetência; oportunidades; trabalhador; ubss	interpretação gerencial e moralizada do desemprego	a rota privilegia desenvolvimento, empregabilidade e serviços
população–desemprego	Online	segurança; homicídios; natalidade; argentina; pobreza; banqueiros	efeitos sociais e agentes econômicos	a rota articula segurança, demografia e economia

Itens opacos, raros ou idiossincráticos

Alguns intermediários exigem cautela interpretativa especial. No par *democracia–ditadura*, obviamente, romantizando e mentalidade são menos transparentes isoladamente do que *arsesp*, *impeachment* ou *massacre*. Em *dinheiro–elite*, realmente, principal, principalmente e exclusivamente têm peso interpretativo menor do que *economia*, *patrimônio* ou *banqueiros*. Em *governo–população*, *questão*, *ainda* e *diferente* são itens de leitura fraca quando considerados isoladamente, ao passo que *gestão*, *polícias* e *homicídios* oferecem relações temáticas mais nítidas.

No par *pobres–elite*, *marginais* e *militares* tensionam a leitura integradora do debate, enquanto *cultural* e *dirigente*, no corpus online, exigem atenção à relação entre dominação econômica e simbólica. Já em *população–desemprego*, *olivato*, *missionária*, *diariamente*, *representa*, *diminuindo* e *menor* são periféricos ou contextualmente opacos. Esses vocábulos permanecem registrados, sem eliminação por conveniência interpretativa. Nem sua persistência prova uma função discursiva específica, nem sua opacidade permite descartá-los como irrelevantes: ambos os aspectos limitam a leitura do núcleo semântico, conforme discutido na Seção 5.6.

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.