INDICADOR DE SIMILARIDADE DO DISCURSO PARLAMENTAR: ANÁLISE DO COMPORTAMENTO DAS COALIZÕES PARTIDÁRIAS

PARLIAMENTARY SPEECH SIMILARITY INDICATOR: ANALYSIS OF BEHAVIOR OF PARTY COALITIONS

INDICADOR DE SIMILITUD DE DISCURSO PARLAMENTARIO: ANÁLISIS DEL COMPORTAMIENTO DE LAS COALICIONES DE PARTIDOS

Fabiano Peruzzo Schwartz[1]

Resumo: Para se compreender a política, faz-se necessário conhecer sobre o que dizem e escrevem os atores políticos. Esse entendimento tem especial significado no sistema político brasileiro em que a organização do Poder Executivo tem como base a formação de grandes coalizões. O presente estudo faz uso do Processamento de Linguagem Natural como ferramenta de análise do discurso parlamentar proferido na Câmara dos Deputados brasileira e propõe indicador específico, baseado na estatística Qui-Quadrado, para a aferição de similaridade de discursos. Os resultados encontrados são coerentes com fatos que marcaram a história política brasileira no período 2001-2015 e revelam que a dimensão ideológica se submete à lógica eleitoral na formação das alianças políticas, sugerindo que o indicador proposto possui potencial explicativo de fenômenos relacionados ao comportamento das coalizões partidárias.

Palavra-chave: Discurso Parlamentar; Processamento de Linguagem Natural; Índice de Identidade Ideológica; Coalizões Partidárias.

Abstract: In order to understand politics, it is necessary to know about what political actors say saying and write. This understanding has special meaning in the Brazilian political system in which the organization of the Executive Power is based on great coalitions. In this sense, the present study makes use of Natural Language Processing as a tool for analyzing parliamentary discourse delivered in the Brazilian Chamber of Deputies and proposes a specific indicator, based on Chi-Square statistics, for the assessment of similarity of discourses. The results found are consistent with facts that marked the Brazilian political history in the period 2001-2015 and reveal that the ideological dimension is subject to the electoral logic in the formation of political alliances, suggesting that the proposed indicator has the potential to explain phenomena related to the behavior of coalitions parties.

Keywords: Parliamentary speech; Natural Language Processing; Ideological Identity Index; Party coalitions.

Resumen: Para entender la política es necesario saber qué dicen y escriben los actores políticos. Esto tiene un significado especial en el sistema político brasileño en el que la organización del Poder Ejecutivo se basa en grandes coaliciones. El presente estudio hace uso del Procesamiento del Lenguaje Natural como herramienta para analizar el discurso parlamentario pronunciado en la Cámara de Diputados de Brasil y propone un indicador específico, basado en la estadística Chi-Cuadrada, para la evaluación de la

[1] Doutor em Engenharia de Sistemas Eletrônicos e de Automação e Mestre em Ciência da Computação, ambos pela Universidade de Brasília. Diretor da Coordenação de Pós-Graduação da Câmara dos Deputados e professor permanente do Mestrado Profissional em Poder Legislativo. Coordena o grupo de pesquisa e extensão "Ciência de Dados Aplicada ao Estudo do Poder Legislativo: abordagem computacional e métodos de análise", cadastrado no Diretório dos Grupos de Pesquisa do CNPq. Orcid Id: https://orcid.org/0000-0003-1727-9346. E-mail: fabiano.schwartz@camara.leg.br

similitud de los discursos. Los resultados encontrados son consistentes con hechos que marcaron la historia política brasileña en el período 2001-2015 y revelan que la dimensión ideológica está sujeta a la lógica electoral en la formación de alianzas políticas, sugiriendo que el indicador propuesto tiene el potencial de explicar fenómenos relacionados al comportamiento de las coaliciones de partidos.

Palabras clave: Discurso Parlamentario; Procesamiento Natural del Lenguaje; Índice de Identidad Ideológica; Coaliciones de Partidos.

1 Introdução

Os trabalhos legislativos e a atividade parlamentar na Câmara dos Deputados são regidos pelo Processo Legislativo[2] e estão intimamente atrelados ao diálogo. No Plenário, órgão máximo de deliberação, os representantes do povo discutem e votam soberanamente as proposições em tramitação. A habilidade retórica de arranjar adequadamente as ideias, de colocar os argumentos apropriados, de organizar e apresentar oralmente o discurso para uma audiência é a principal ferramenta do parlamentar no exercício do seu ofício. Gnerre (1991, p. 5) destaca que "o poder da palavra é o poder de mobilizar a autoridade acumulada pelo falante e concentrá-la num ato linguístico", do qual um dos mais eloquentes exemplos é o discurso político.

Perceber a necessidade do discurso parlamentar e a respectiva influência no Processo Legislativo, do qual resultarão as leis que vão atingir direta ou indiretamente a vida de todos os cidadãos brasileiros, é de vital importância. Ademais, entender a dinâmica da política em espectro amplo somente é possível quando se conhece sobre o que os atores políticos estão dizendo e escrevendo, como se posicionam ante os fatos que afetam a sociedade. Uma vez que o discurso é um modo de representar aspectos do mundo em um contexto específico, a articulação de mais de um deles é a chave para reunir diferentes perspectivas de compreensão da realidade. Portanto, a análise da atuação parlamentar por meio do discurso deve ser considerada pelo conjunto da obra, não apenas por pronunciamentos isolados, pois esses carregam o peso do momento e do ambiente em que foram proferidos. Há que se considerar a sequência de eventos para se identificar a evolução do discurso e desta extrair o conhecimento sobre o processo político.

Esse entendimento tem especial significado no sistema político brasileiro em que a organização do Poder Executivo tem como base a formação de grandes coalizões partidárias, ao que fora denominado ‘presidencialismo de coalizão’ (ABRANCHES, 1988). Na prática, um governo precisa formar base de apoio no Congresso Nacional para viabilizar suas iniciativas de implantação da política estatal. Portanto, acompanhar o comportamento dessa base nas discussões das diversas matérias apreciadas no parlamento, o que significa acompanhar o comportamento dos parlamentares que a compõem, é parte intrínseca ao jogo político. Tanto

[2] Processo Legislativo é o conjunto de atos realizados pelos órgãos do Poder Legislativo, seguindo regras fixadas para a elaboração de normas jurídicas como emendas à Constituição, leis complementares ou ordinárias e outros tipos normativos dispostos no art. 59 da Constituição Federal (BRASIL, 1988).

que a Constituição Federal de 1988, em seu artigo 17, § 1º, ressalta que os partidos políticos têm autonomia para "adotar os critérios de escolha e o regime de suas coligações nas eleições majoritárias" e que devem estabelecer em seus estatutos "normas de disciplina e fidelidade partidária" (BRASIL, 1988).

Não há no texto constitucional definição expressa dos termos "disciplina" e "fidelidade" partidárias, o que leva a divergências doutrinárias quanto a sua interpretação e aplicação. A linha da doutrina que diferencia os termos, atrela: fidelidade partidária ao alinhamento entre parlamentar e partido político no âmbito ideológico, ou caráter filosófico-programático dos temas; disciplina partidária ao comportamento parlamentar frente a questões cotidianas do partido, observada via o cotejamento entre os votos nominais dos parlamentares e as orientações dos líderes de bancada. Nesse sentido, Roma (2007) categoriza dois blocos de parlamentares: fiéis ou infiéis, quando o parlamentar, após eleito, permanece no partido, ou migra para outro; e disciplinados ou não disciplinados, quando os votos no plenário da Câmara seguem, ou não, a orientação dos líderes dos partidos.

Portanto, espera-se que as coalizões sejam marcadas pela fidelidade e disciplina partidárias dos seus membros. Nessa vertente que distingue os termos, a aferição da disciplina tem caráter objetivo, bem definido, que permite quantificar por meio das votações em dado período o percentual de vezes que o parlamentar ou a bancada seguem a orientação do líder. Assim, o partido pode monitorar o índice de disciplina dos seus parlamentares ao longo do tempo. Já a fidelidade, ou melhor, a infidelidade é medida em fato único, quando o parlamentar deserta ou abandona o partido, migrando de legenda. A identificação de sinais de infidelidade ao longo do tempo é um processo subjetivo, que deve observar outros aspectos do comportamento parlamentar.

Nessa perspectiva, o discurso é matéria prima fértil na medida em que, por definição, carrega traço ideológico. Segundo Orlandi (2015, p. 43), as formações discursivas, que determinam o que deve ser dito em dada conjuntura sócio-histórica, representam no discurso as formações ideológicas, não havendo, nas palavras, sentido que não seja determinado ideologicamente. Há, assim, uma relação recíproca entre ideologia e linguagem.

Também existe relação entre ideologia e a lógica das coalizões partidárias, no que Machado e Miguel (2011) descrevem como dimensão programática "vinculada aos valores políticos de base, que leva em conta a formação de alianças entre partidos [...] segundo a sua categorização por ideologia". Destacam que uma coligação é tão mais coerente quanto maior a afinidade das posições ideológicas de seus membros.

Logo, é razoável considerar que do discurso parlamentar podem ser extraídas posições ideológicas e que a verificação sistemática dessas posições é capaz de revelar afinidades ideológicas. No âmbito das coalizões partidárias é também razoável associar tais afinidades a um grau de coerência, ou identidade. E porque não dizer um grau de fidelidade. Ou seja, a

fidelidade partidária não seria apenas caracterizada pelo ato da deserção, mas poderia ser mensurada e acompanhada ao longo do tempo por aquilo que os parlamentares dizem em pronunciamento, podendo ser o monitoramento no indivíduo ou num grupo de indivíduos.

O principal problema reside na dificuldade de se efetuar essa verificação sistemática, uma vez que exige grande esforço de levantamento dos dados (textos dos discursos) e de análise. Análise que não se restringe à de conteúdo, pois não procura somente extrair sentido do texto discursivo, mas compreendê-lo de forma conjugada à história dos acontecimentos. Tomando por base o resgate da linha argumentativa a partir dos registros taquigráficos dos discursos efetuados no Plenário da Câmara dos Deputados, a execução manual do processo de verificação seria morosa e imprecisa, além de não conseguir contemplar a totalidade dos pronunciamentos.

Alternativamente, a captação e a gestão do conhecimento implícito no pronunciamento parlamentar podem ser aprimoradas pelo uso de recursos computacionais e técnicas avançadas de processamento linguístico. Cientistas políticos têm usado a análise automática de conteúdo em um conjunto diversificado de textos. Isso inclui arquivos de dados de mídia (YOUNG; SOROKA, 2012); discursos parlamentares em legislaturas de todo o mundo (QUINN *et al*., 2010); declarações do presidente, do legislador e do partido (GRIMMER, 2010); tratados (SPIRLING, 2012); artigos de ciência política e outros textos políticos.

Nesse sentido, o presente estudo faz uso das técnicas do Processamento de Linguagem Natural (PLN) como ferramenta de tratamento e auxílio à análise do discurso parlamentar, e como recurso útil no estabelecimento de conexões entre fatos, dados quantitativos e resultados finais do Processo Legislativo. A partir do preparo dos dados por meio do PLN, esta pesquisa inova ao propor um indicador de similaridade textual, ora denominado Índice de Identidade Ideológica, com base na distribuição Qui-Quadrado (χ^2), para se aferir o grau de convergência do pronunciamento de deputados pertencentes a determinado bloco parlamentar ou coalizão. Essa aferição permite investigar, em certa medida, a solidez de alianças políticas ou, por extensão, a fidelidade partidária.

O artigo está divido em sete seções, além desta Introdução: a segunda seção discute o conceito de ideologia político-partidária e a formação de coalizões; a terceira, apresenta técnicas do Processamento de Linguagem Natural; a quarta, desenvolve a matemática para se aferir a similaridade do discurso; a quinta, propõe o Índice de Identidade Ideológica; a sexta, discute os resultados e principais achados; a sétima, traz as considerações finais.

2 Ideologia Político-Partidária e Coalizão

Aqui não há a pretensão de se discutir o conceito "ideologia" em sua amplitude teórico-filosófica, mas, tão somente, de se estabelecer um referencial pragmático, um recurso metodológico com o propósito de identificar eventos, a partir do discurso parlamentar, que

possam revelar alguma forma de afinidade, lealdade partidária e/ou coerência do discurso, capaz de caracterizar uma coalizão.

A primeira teoria sobre o processo de formação de coalizões, desenvolvida pelo cientista político William H. Riker, não pressupunha proximidade ideológica (RIKER, 1962). Talvez pelo fato de o termo ideologia ser compreendido por diversos significados, ora divergentes, essa associação somente tenha sido proposta anos mais tarde por Robert Axelrod (AXELROD, 1970).

Argumentando que uma definição única de ideologia seria inútil, Eagleton (1997) apresentou uma lista das mais comuns, donde se destacam: conjunto de crenças orientadas para a ação; corpo de ideias característico de um determinado grupo ou classe social; ideias que ajudam a legitimar um poder político dominante; pensamento de identidade; a conjuntura de discurso e poder. Segundo Cancian (2007), em pesquisas empíricas, o termo "ideologia" é usado com o objetivo de "descrever o conjunto de ideias, valores ou crenças que orientam a percepção e o comportamento dos indivíduos sobre diversos assuntos ou aspectos sociais", como, por exemplo, "as opiniões e as preferências que os indivíduos têm sobre o sistema político vigente, a ordem pública, o governo, as leis, as condições econômicas e sociais". Para ambos os autores, prevalece nos conceitos o sentido de unidade, de orientação comum ou identidade, que também tem relação com o significado de coligar.

Machado e Miguel (2011), ao proporem uma tipologia para coligações partidárias, simplificaram o uso do conceito ideologia assumindo que "despida dos seus significados mais complexos (e mais polêmicos)" a palavra ideologia "remete unicamente à posição no eixo esquerda-direita". Rodrigues (2009, p. 27), em seu estudo sobre bancadas partidárias, discute a lógica ideológica das coligações sob duas vertentes: por um lado, acredita-se que as coligações recebem avaliação negativa da opinião pública por unirem legendas ideológica e programaticamente discrepantes, num cenário de acentuada migração partidária; por outro, defende-se que a maioria das coligações obedece à lógica da afinidade ideológica.

Reisman (2016) destaca que a formação das coalizões políticas está localizada temporalmente no momento pré-eleitoral, quando se negociam alianças estratégicas entre partidos políticos para as eleições majoritárias e proporcionais. Ao estudar as coalizões eleitorais lideradas pelo PT em campanhas presidenciais, o autor constatou que, ainda que os critérios de formação pela ideologia pareçam ser dominantes, essa suposta convergência ideológica não impediu a movimentação das bandeiras de coalizões da esquerda para o centro, mostrando que a ideologia se submete à lógica eleitoral. Tais coalizões, ao longo do tempo, passaram por drástica redução na carga ideológica e adotaram programas eleitorais genéricos e abstratos. Em sua vertente pragmática, essas coalizões abandonaram defesas de pontos polêmicos e aproximaram-se do centro do eixo esquerda-direita.

O fenômeno observado no estudo de Reisman (2016), presente no cenário político geral,

é provavelmente o fator motivador da Emenda Constitucional n. 97, de 4 de outubro de 2017, que vedou as coligações partidárias nas eleições proporcionais (BRASIL, 2017), talvez para afastar a lógica eleitoral do processo de formação das coalizões, estimulando a convergência ideológica na concepção dos programas.

Portanto, neste estudo se propõe, de forma simplificada, a utilização do conceito "ideologia" com o complemento "político-partidária", representando o conjunto de orientações políticas com o qual um partido se compromete. Ainda que esse conjunto sofra alterações com o tempo, deve refletir o compromisso do partido num dado contexto temporal. À luz dessa denominação, "ideologia político-partidária", e sob a perspectiva da formação das coalizões partidárias, poder-se-ia esperar que essas associações ocorressem, por princípio, entre partidos que compartilhassem das mesmas orientações políticas, ou, ao menos, da maior parte delas. Assumindo essa hipótese como bastante provável, poderíamos supor que dois partidos coligados pertenceriam, ou deveriam pertencer, à mesma ideologia político-partidária.

Com base nesses aspectos, duas premissas podem ser estabelecidas para a condução do presente estudo:

1. por princípio, uma coalizão deve ser fundada no conceito proposto de ideologia político-partidária; logo, espera-se alto grau de similaridade entre os discursos de parlamentares de partidos coligados;

2. um indicador para medir a identidade ideológica de uma coalizão deve ser capaz de captar o comportamento de todo um grupo de partidos em períodos determinados.

Logo, o principal desafio consiste em medir o grau de similaridade entre discursos dos diversos parlamentares de uma coalizão, o que requer o processamento de conteúdo de textos em larga escala, na ordem de centenas ou milhares, capacidade somente possível por meio de recursos computacionais.

As duas próximas seções descrevem os ferramentais técnico-metodológicos utilizados neste estudo para a construção de um indicador de identidade ideológica. São apresentados a técnica do saco-de-palavras e o método estatístico Qui-Quadrado com os seguintes propósitos: preparar as coleções de texto para o tratamento computacional; verificar se agrupamentos específicos de duas palavras compõem um significado próprio da língua; e estabelecer métrica de similaridade entre coleções de textos. São utilizados o conceito de *corpora*, para se referir a uma grande coleção de textos digitais sistematicamente organizada sobre a qual uma análise linguística é baseada, e o de *corpus* (singular de *corpora*), para um subconjunto de textos extraídos do *corpora* (VYATKINA; BOULTON, 2017).

3 Processamento de Linguagem Natural

O Processamento de Linguagem Natural é uma subárea da Inteligência Artificial e da Linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais. Grimmer e Stewart (2013), em estudo de textos políticos, reconhecem os benefícios do PLN e o quanto o processamento automatizado reduz custos e esforços de análise de grandes coleções de texto. Ressaltam, contudo, que métodos automatizados não substituem um pensamento cuidadoso e uma leitura atenta, além de exigirem uma validação extensa e específica do problema. Argumentam que, para que o processamento automatizado de texto se torne uma ferramenta padrão para cientistas políticos, os pesquisadores da área devem contribuir com a criação de formas robustas de validação dos métodos.

Zhang *et al*. (2009) afirmam que a capacidade de expressar o sentido de uma palavra depende das demais palavras que a acompanham. Quando uma palavra aparece acompanhada por um conjunto de termos, maiores são as chances desse conjunto possuir um significado relevante. Isso quer dizer que não apenas a palavra, mas também a informação contextual é útil para o processamento de informações.

Silva e Souza (2014) destacam que o texto não é um simples amontoado aleatório de palavras, mas que a ordem em que são colocadas é que produz o significado. O estudo da co-ocorrência das palavras pode indicar se estas estão relacionadas diretamente, por composicionalidade ou afinidade, ou indiretamente, por semelhança. Portanto, a base da linguística empírica consiste em encontrar, a partir da frequência de co-ocorrências observadas, as dependências significativas entre os termos. Esses termos adjacentes são denominados n-gramas, onde *n* é a quantidade de termos.

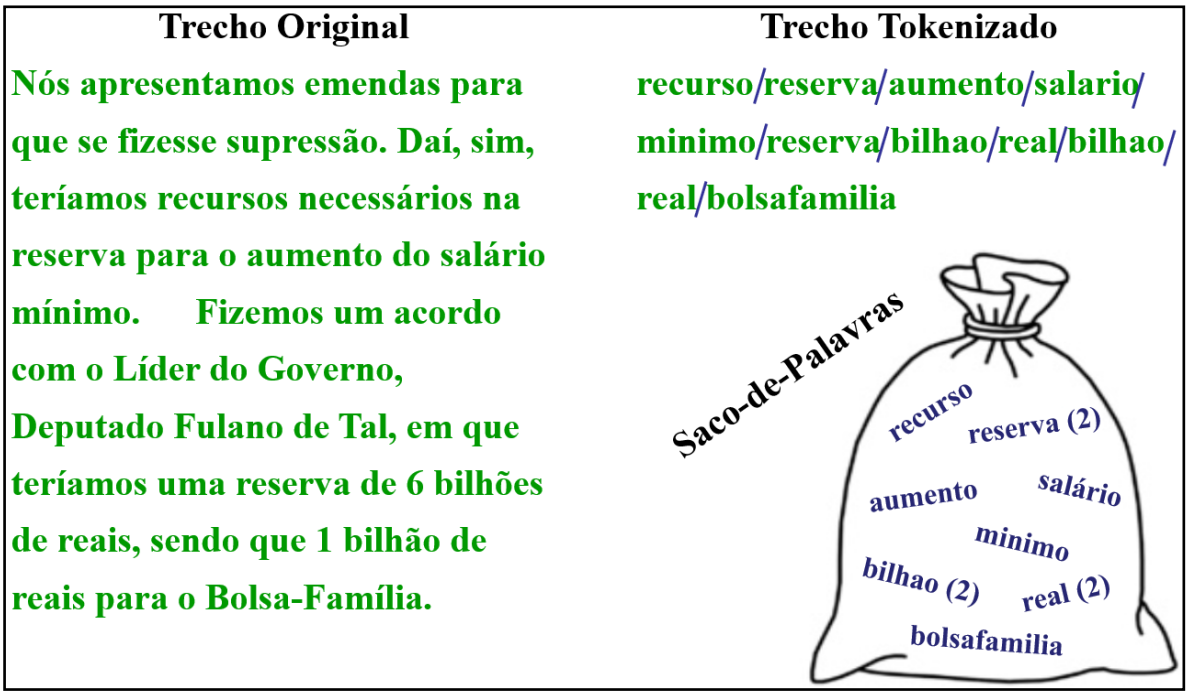
Os modelos de linguagem n-gramas foram desenvolvidos no âmbito da Estatística pelo matemático russo Andrey Markov (1856-1922) com a finalidade de reconhecer padrões estatísticos do uso da língua baseados em cadeias, conhecidas como cadeias de Markov. Wang e Liu (2011) apontam que muitos trabalhos têm como foco dominante a identificação e extração de n-gramas. Esse processo é denominado tokenização ou segmentação de palavras e consiste na tarefa de dividir um texto em unidades mínimas chamadas tokens, onde cada token pode ser uma palavra ou conjunto de palavras, um número, um sinal de pontuação ou outra estrutura pertencente à linguagem (MANNING; SCHÜTZE, 1999).

A complexidade da tokenização varia de acordo com a complexidade do próprio idioma, com a definição do problema a ser estudado e com o modelo escolhido para aplicação do PLN. O processo de tokenização geralmente inclui as seguintes etapas: conversão dos caracteres do texto para minúsculo; remoção de números, pontuação, plurais, sinais de acentuação, *stopwords*[3] e espaços em branco; e segmentação das palavras. O produto final da

[3] *Stopwords* são palavras com pouco significado em algumas aplicações de PLN, como a recuperação e classificação

tokenização é o saco-de-palavras (do inglês, *bag-of-words*) com a respectiva contagem de frequência dos termos. Na abordagem do saco-de-palavras, a ordem em que as palavras estão dispostas no texto não é considerada e a busca por informações leva em conta apenas as frequências estimadas. A Figura 1 ilustra um trecho de discurso antes e depois do processo de tokenização. A redução significativa do texto se deve, principalmente, ao conjunto de *stopwords* utilizado.

Figura 1 – Tokenização aplicada a trecho de discurso parlamentar: saco-de-palavras.



Fonte: Elaboração própria.

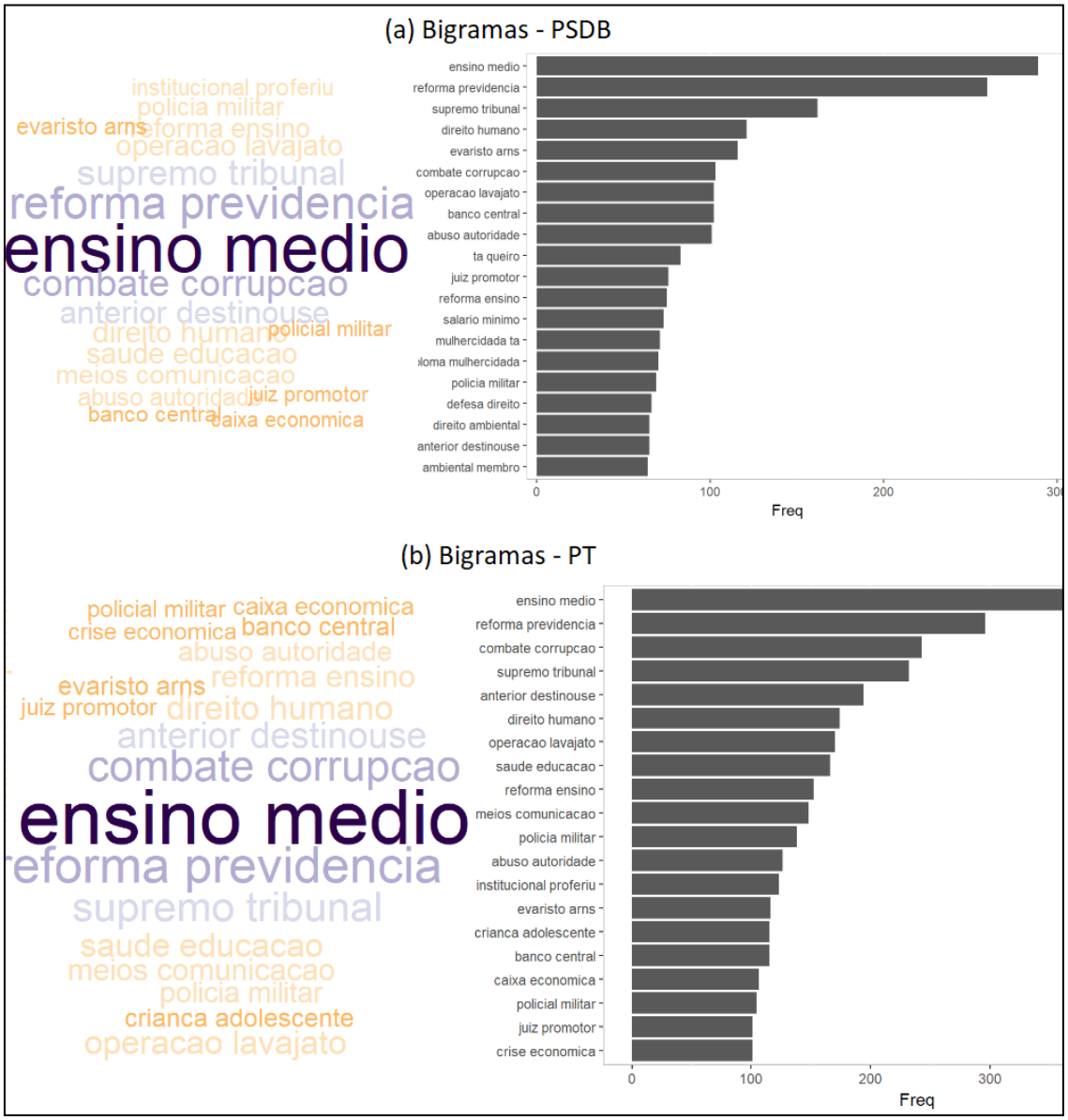
O processo de tokenização pode ser ajustado para efetuar a divisão do texto em bigramas, ou seja, tokens com duas palavras que ocorrem consecutivamente no texto. Os bigramas são comumente utilizados como base para a análise estatística de textos e ajudam a estimar a probabilidade condicional de uma palavra, dada a palavra precedente (JURAFSKY; MARTIN, 2008). Bigramas são mais ricos em significado do que a apuração de palavras simples, sendo mais úteis na extração de informações do texto.

A Figura 2 ilustra o conjunto de bigramas extraídos dos discursos proferidos pelos deputados do PSDB e do PT no ano de 2016, apresentados nas formas de nuvens de palavras e de gráficos de frequência. Esses recursos constituem um método heurístico de análise que, por si só, não permite generalizar a resposta a uma questão de pesquisa, mas aponta caminhos para

de informações, o que significa que essas palavras não são muito discriminatórias. Geralmente, artigos e pronomes são classificados como *stopwords*. A ideia é simplesmente remover as palavras que ocorrem com muita frequência em todos os documentos da coleção de textos. Contudo, a escolha da lista de *stopwords* deve ser cuidadosa pois é determinante para o alcance dos objetivos.

o que deve ou não receber maior atenção em um conjunto de textos.

Figura 2 – Nuvem de palavras e gráfico de barras de frequências absolutas: (a) bigramas dos discursos do PSDB em 2016; (b) bigramas dos discursos do PT em 2016.



Fonte: Elaboração própria.

Num primeiro olhar sobre a Figura 2, contrariando o que o senso comum poderia esperar sobre pronunciamentos de partidos declaradamente opostos, tem-se a impressão de que os discursos são semelhantes e abordam as mesmas temáticas, com pequenas variações. Contudo, embora importante para o amadurecimento do processo investigatório, a inspeção visual não é suficiente para se avançar em conclusões, sendo necessárias abordagens mais robustas por meio de testes estatísticos, que serão detalhados na próxima seção.

4 Similaridade do discurso por comparação de bigramas como *collocations*

A tarefa de mensurar o quanto um *corpus* de discursos se assemelha a outro por meio de método computacional deve preservar, no que for possível, os significados dos textos. Muito embora a técnica do saco de palavras não tenha como função a avaliação semântica, visto que desconsidera a ordem em que as palavras estão dispostas no texto, a divisão em n-gramas com n maior do que 1 ($n > 1$) resulta em conjuntos de termos cuja ordem corresponde à do texto original. Portanto, esses n-gramas carregam mais informação semântica do que termos isolados, sendo razoável supor, por exemplo, que a comparação entre discursos com base em bigramas é mais eficaz do que aquela baseada em termos isolados.

Essa suposição ganha ainda mais força se formos capazes de identificar os bigramas cujos termos não acontecem juntos por acaso, mas por alguma razão de dependência que expresse uma forma convencional, na linguagem, de dizer as coisas, isto é, que esteja inserida em um contexto semântico. Expressões de duas ou mais palavras que correspondam a uma forma convencional de dizer as coisas são chamadas de colocações, ou *collocations* (MANNING; SCHÜTZE, 1999).

O teste estatístico Qui-Quadrado (X^2) pode ser utilizado para avaliar se duas palavras constituem uma *collocation*, verificando a dependência entre elas com o auxílio de uma tabela de contingência (KILGARRIFF; ROSE,1998). Em essência, o teste trata cada palavra como uma variável categórica e assume a condição de independência dos eventos, ou seja, quando as palavras acontecem juntas ao acaso e, portanto, não constituem uma *collocation*. Por definição, dois eventos são independentes quando a probabilidade de ambos acontecerem juntos é igual ao produto das probabilidades de cada evento acontecer individualmente:

$$P(AB) = P(A)P(B) \tag{1}$$

Então, o valor esperado para o acontecimento conjunto de dois eventos independentes é dado por

$$E(AB) = P(A)P(B) * N \tag{2}$$

onde N é o número total de eventos.

Assim, conhecendo-se o valor esperado dos eventos independentes, a estatística X^2 é determinada comparando-se as frequências observadas com as frequências esperadas. Quando não há diferença estatística entre as frequências observadas e as frequências esperadas, conclui-se que os eventos são independentes. As hipóteses sob a condição de independência dos termos do bigrama são enunciadas por:

H_0: Não há associação entre os termos; o bigrama não é uma *collocation*.

H_1: Há associação entre os termos; o bigrama é uma *collocation*.

Na prática, é computada a frequência com que as palavras do bigrama ocorrem conjuntamente, as frequências com que formam bigramas com outras palavras e a frequência de bigramas formados sem essas palavras. A Tabela 1 mostra a distribuição dos termos “ensino” e “medio” (sem acento após a tokenização) para a análise do bigrama “ensino medio", onde w_1 e w_2 se referem, respectivamente, à primeira e à segunda palavra do bigrama.

Tabela 1 – Tabela de contingência: distribuição dos termos “ensino” e “medio” para análise do bigrama “ensino medio”.

	w_1 = ensino	$w_1 \neq$ ensino
w_2 = medio	478 (ensino medio)	53 (p.e., oriente medio)
$w_2 \neq$ medio	269 (p.e., ensino superior)	32.638

Nota: frequências aferidas a partir dos discursos do PT em 2016.

Fonte: Elaboração própria.

Então, a partir da tabela de contingência, estima-se a estatística X^2 da seguinte forma:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

onde i representa as linhas e j as colunas da tabela, O_{ij} é o valor observado na célula (i, j) e E_{ij} é o valor esperado. Os valores de E_{ij} são computados conforme a Equação 2 a partir das probabilidades marginais, isto é, a partir dos totais de linhas e colunas convertidos em proporções. Por exemplo, para a célula (1, 1), o valor esperado E_{11} é dado pelo produto entre a probabilidade de “ensino” ser o primeiro termo do bigrama ($P(A)$), a probabilidade de “medio” ser o segundo termo do bigrama ($P(B)$) e a quantidade total de bigramas (N) existentes no *corpus* em análise:

$$E_{11} = \frac{478+269}{33.438} * \frac{478+53}{33.438} * 33.438 \approx 11{,}86 \quad (4)$$

Aplicando-se a Equação 3 a todas as células da Tabela 1, encontramos $X^2 \approx 28{,}75$ e $p \approx 0.0000000821$, o que significa que H_0 deve ser rejeitada, ou seja, as palavras “ensino” e “medio” não ocorrem juntas ao acaso. São, portanto, dependentes e constituem uma *collocation*.

Uma vez identificadas as *collocations* de dois *corpus*, a similaridade entre eles pode ser medida, mais uma vez, por meio da estatística X^2. Kilgarriff e Rose (1998) propuseram essa abordagem em estudo onde compilaram uma tabela de n linhas por 2 colunas (com n = 500), na qual cada coluna correspondia à contagem das palavras mais frequentes comuns aos dois *corpus*. No presente estudo, o método de Kilgarriff e Rose (1998) foi estendido para bigramas

validados como *collocations*. A Tabela 2 contém uma representação esquemática dessa concepção, utilizando os *corpus* de discursos do PT e PSDB em 2016.

Tabela 2 – *Collocations* mais frequentes comuns aos *corpus* de discursos do PT e PSDB em 2016.

	collocation	**freq. PT**	**freq. PSDB**
1	ensino medio	478	289
2	reforma previdencia	296	260
3	combate corrupcao	243	103
4	supremo tribunal	232	162
5	direito humano	174	121
6	operacao lavajato	170	102
7	saude educacao	166	64
8	reforma ensino	152	75
9	meios comunicacao	148	55
10	policia militar	138	60
...			
n	....	....	....

Fonte: Elaboração própria.

Para cada uma das *n collocations* mais frequentes da Tabela 2 é necessário calcular o número de ocorrências esperadas (valor esperado) em cada *corpus*, no intuito de verificar se ambos podem ser considerados amostras aleatórias de uma mesma população. Dados os tamanhos N_1 e N_2, dos *corpus* 1 e 2, e as frequências observadas O_{i1} e O_{i2} da *i*-ésima *collocation* nos respectivos *corpus*, então, o valor esperado para essa *collocation* no *corpus* 1 é dado pela probabilidade de a *collocation* acontecer no conjunto total (*corpus* 1 mais *corpus* 2) multiplicada pelo tamanho do *corpus* 1 (N_1)

$$E_{i1} = \frac{(O_{i1}+O_{i2})}{N_1+N_2} * N_1 \tag{5}$$

e, respectivamente no *corpus* 2, por

$$E_{i2} = \frac{(O_{i1}+O_{i2})}{N_1+N_2} * N_2 \tag{6}$$

Conhecendo-se os valores observados e esperados, então X^2 pode ser determinado pela Equação 7,

$$X^2 = \sum_{c=1}^{2}\sum_{i=1}^{n}\frac{(O_{ic}-E_{ic})^2}{E_{ic}} \tag{7}$$

onde *c* representa o respectivo *corpus*. As hipóteses nula e alternativa podem ser formuladas como segue:

H_0: As frequências observadas não diferem das frequências esperadas, isto é, não existe diferença entre as frequências dos grupos.

H_1: As frequências observadas são diferentes das frequências esperadas, portanto, existe diferença entre as frequências dos grupos.

Portanto, aceitar H_0 significa dizer que não existe diferença entre os *corpus*.

Definidos os ferramentais computacional e estatístico, a seção seguinte aplica os conceitos explorados na construção do indicador de similaridade proposto, baseado no conceito de ideologia político-partidária.

5 Índice de Identidade Ideológica

A partir da proposta de Kilgarriff e Rose (1998) estendida para bigramas, um primeiro teste de similaridade foi efetuado para combinações dos partidos PT, PSDB, PMDB[4], PSOL, PCDOB e PTB, cujos *corpus* do ano de 2016 foram comparados dois a dois por meio das respectivas *collocations* (Tabela 3).

Tabela 3 – Métrica X^2 para similaridade de corpus: partidos PT, PSDB, PMDB, PSOL, PCdoB e PTB; ano 2016.

Partido 1	**Partido 2**	**Bigrama (*collocation*)**		
		X^2	***p***	**H_0**
PT	PSDB	2.386,57	7,24e-302	rejeita
PT	PMDB	1.836,77	1,67e-196	rejeita
PT	PSOL	4.093,27	0,00	rejeita
PT	PCdoB	4.217,40	0,00	rejeita
PT	PTB	4.390,26	0,00	rejeita
PSDB	PMDB	165,68	1,00	aceita
PSDB	PSOL	1.679,38	1,91e-221	rejeita
PSDB	PCdoB	1.763,91	2,38e-238	rejeita
PSDB	PTB	3.289,55	0,00	rejeita
PMDB	PSOL	1.919,91	7,75e-277	rejeita
PMDB	PCdoB	1.994,99	1,83e-292	rejeita
PMDB	PTB	3.113,44	0,00	rejeita
PSOL	PCdoB	16,83	1,00	aceita
PSOL	PTB	708,59	8,29e-32	rejeita
PCdoB	PTB	648,03	8,91e-23	rejeita

Fonte: Elaboração própria.

Observa-se na Tabela 3 a similaridade de *corpus* nas combinações PSDB/PMDB e

[4] No período correspondente à estruturação do *corpora*, o partido ainda atendia pela sigla PMDB, alterada, posteriormente, para MDB.

PSOL/PCdoB. Em análise preliminar, isso parece razoável num ano em que o *impeachment* da presidenta Dilma dominou o cenário político, no qual partidos como PSDB e PMDB se aliaram à causa, ao passo que outros, como PSOL e PCdoB, posicionaram-se contrariamente. Observa-se, também, que o discurso do PT não se assemelha ao de nenhum outro partido, talvez refletindo uma quebra da base governista diante dos inúmeros interesses que circundaram o tema *impeachment*.

Esse primeiro experimento se mostrou coerente na comparação dos *corpus*, revelando capacidade explicativa dos acontecimentos, o que inspirou a construção de indicador com a função de mensurar o grau de similaridade do discurso de um grupo de partidos afins, denominado **Índice de Identidade Ideológica**. A ideia consiste em **determinar a probabilidade média das estatísticas X^2 encontradas na comparação dos discursos de partidos pertencentes a um dado grupo, tomados dois-a-dois**. Pelo pressuposto assumido da ideologia político-partidária, deve-se esperar que esse indicador tenha valor muito próximo de 1 (um) quando mensurado sobre partidos de uma mesma coligação. Por analogia, quando aplicado a partidos com orientações políticas opostas, deve produzir valores próximos de zero.

Para fins deste estudo, foram considerados os discursos proferidos no Pequeno[5] e Grande Expediente[6]. A escolha dessas fases da sessão ordinária do Plenário se deve ao fato de que favorecem um debate de caráter ideológico, menos influenciado pela pressão das votações e de humores políticos acentuados.

Sobre isso, Moreira (2016) identificou que os discursos proferidos no Pequeno Expediente não apresentam concentração de temas, "havendo indícios de que tal estratégia é orientada especialmente pela quantidade de discursos proferidos e pela ideologia da legenda". Acrescenta o autor que "há entre os parlamentares diferenças de foco temático" e que "tal diferença sofre a influência do percentual de votos recebidos, da ideologia da legenda partidária pela qual o parlamentar foi eleito".

Esse fenômeno pode ser decorrente da liberdade que o parlamentar tem durante o Pequeno Expediente, podendo discursar sobre assuntos livres por até 5 (cinco) minutos. Ainda segundo Moreira (2016), "muitos deputados passam o mandato inteiro sem falar nas votações e comissões, mas no Pequeno Expediente quase todos discursam", mostrando que, nas legislaturas de número 51 a 54, 88% dos deputados realizaram ao menos um discurso no Pequeno Expediente, com uma média de 59 falas por deputado. Da mesma forma, o Grande Expediente foi incluído por guardar duas semelhanças com o Pequeno Expediente: o deputado manifesta a intenção de discursar por meio de inscrição; existe liberdade para o tema do discurso.

[5] Primeira parte da sessão ordinária do Plenário, tem duração máxima de 60 minutos e é destinado às comunicações de parlamentares previamente inscritos (CÂMARA DOS DEPUTADOS, 2020, p. 48).

[6] Fase da sessão plenária que sucede à do Pequeno Expediente com duração improrrogável de cinquenta minutos. Destina-se ao pronunciamento de parlamentares inscritos por até vinte e cinco minutos para cada orador, incluídos aí os eventuais apartes concedidos (CÂMARA DOS DEPUTADOS, 2020, p. 49-50).

6 Resultados e discussão

Foram efetuadas três medições distintas do Índice de Identidade Ideológica, ano-a-ano, no período de 2001 a 2015: a primeira considera somente os partidos pertencentes à coalizão de governo (ou base governista); a segunda considera somente os partidos não pertencentes à coalizão de governo (aqui definido, de forma geral, como oposição); a terceira, estima a similaridade do discurso entre partidos pertencentes e os não pertencentes à coalizão de governo. Os dados sobre as composições das coalizões de governo a cada ano foram extraídos da Base de Dados Legislativos do Centro Brasileiro de Análise e Planejamento – Cebrap (CEBRAP, 2019). O inteiro teor dos discursos proferidos no Plenário da Câmara, no referido período, foi baixado do Portal de Dados Abertos da Câmara dos Deputados (CÂMARA DOS DEPUTADOS, 2017). No total, foram lidos e processados mais de 150 mil discursos, conforme dados e códigos de programação na Linguagem R disponíveis em repositório público (CEFOR, 2018).

Para cada ano do período foi gerada uma matriz de similaridade como a da Tabela 4, que retrata os valores estimados para o ano de 2003. Os partidos grafados em caixa alta correspondem à coalizão governista.

Tabela 4 – Matriz de similaridade do discurso para o ano de 2003: $\mathbf{i_{coa}}$ é o índice de identidade ideológica para a coalizão governista; $\mathbf{i_{opo}}$ é o índice de identidade ideológica para a oposição; $\mathbf{i_{dif}}$ é o índice de identidade ideológica quando grupos diferentes são comparados.

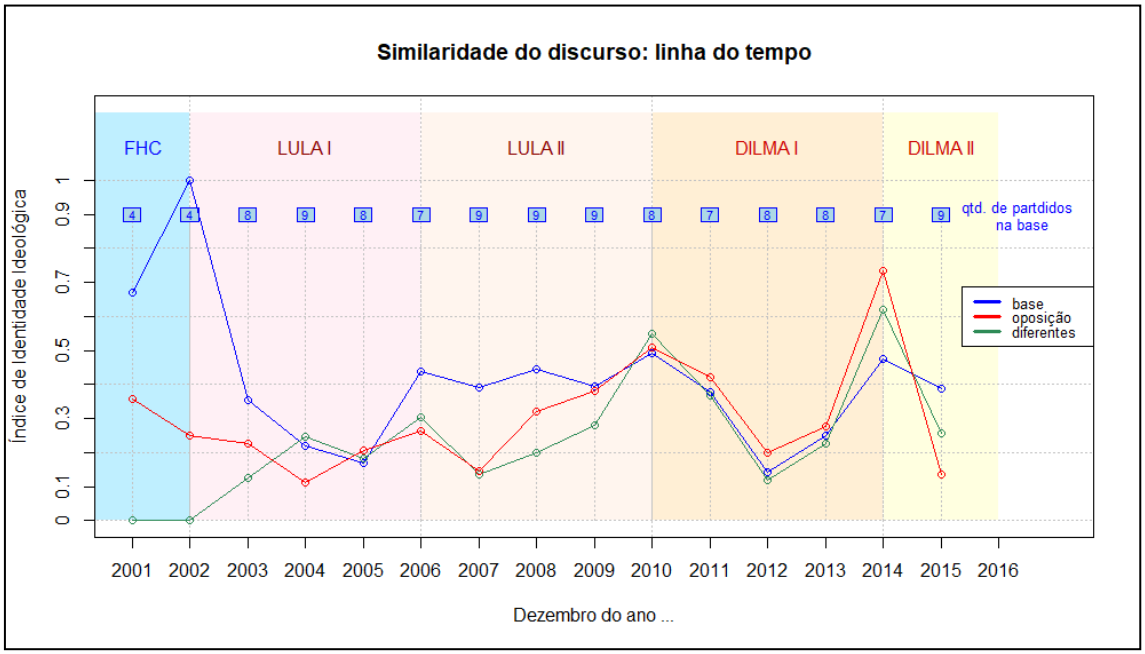
	pfl	PCdoB	PDT	pmdb	pp	PPS	PR	prona	PSB	psc	psdb	PT	PTB	PV	i_{coa}	i_{opo}	i_{dif}
pfl	1,00	0,00	0,00	0,97	0,00	0,00	0,00	0,00	0,00	0,00	0,68	0,00	0,00	0,00	–	0,33	0,00
PCdoB	0,00	1,00	1,00	0,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,43	–	0,17
PDT	0,00	1,00	1,00	0,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,43	–	0,17
pmdb	0,97	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	–	0,39	0,00
pp	0,00	1,00	1,00	0,00	1,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	–	0,00	0,62
PPS	0,00	1,00	1,00	0,00	1,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,57	–	0,17
PR	0,00	0,00	0,00	0,00	1,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	1,00	0,00	0,43	–	0,17
prona	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,74	0,00	0,00	0,00	1,00	–	0,15	0,13
PSB	0,00	0,00	0,00	0,00	1,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	0,95	0,00	0,42	–	0,17
psc	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,74	0,00	1,00	0,00	0,00	0,00	0,00	–	0,15	0,00
psdb	0,68	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	–	0,34	0,00
PT	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	–	0,00
PTB	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,95	0,00	0,00	0,00	1,00	0,00	0,28	–	0,00
PV	0,00	1,00	1,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	1,00	0,29	–	0,17
														Média	**0,36**	**0,23**	**0,12**

Fonte: Elaboração própria.

Nota: os partidos grafados em caixa alta correspondem à coalizão governista; as células de cruzamento dos partidos representam a probabilidade da estatística X^2 resultante da comparação dos respectivos partidos; são destacadas duas linhas como exemplo, uma para a coalizão (PCdoB) e outra para a oposição (pmdb); as células em azul destacam os termos utilizados na respectiva linha para a estimativa do $\mathbf{i_{coa}}$; as células em salmão destacam os termos utilizados na respectiva linha para a estimativa do $\mathbf{i_{opo}}$; as células na cor verde destacam os termos utilizados na respectiva linha para a estimativa do $\mathbf{i_{dif}}$.

A Figura 3 ilustra os resultados encontrados para o indicador no período de 2001 a 2015, destacando os respectivos governos da época.

Figura 3 – Índice de Identidade Ideológica no período de 2001 a 2015.



Fonte: Elaboração própria.

É possível identificar relação coerente entre os resultados da Figura 3 e fatos ocorridos em cada fase de governo, de forma a se avaliar o potencial analítico do indicador proposto. Um primeiro padrão observado é o da ascendência do indicador da base governista nos anos eleitorais (2002, 2006, 2010 e 2014), seguida por queda no ano posterior ao da eleição. Esse comportamento tende a confirmar a argumentação de Reisman (2016) de que a ideologia se submete à lógica eleitoral. Movimento semelhante é observado no indicador de oposição, exceto para o ano 2002, ano em que Lula seria eleito para o seu primeiro mandato à Presidência da República. Nessa ocasião, quando Lula começou a crescer nas pesquisas de opinião de voto, um clima de insegurança se espalhou nas áreas política e econômica, sob a desconfiança de que o país poderia falir. Lula se viu obrigado a assinar texto que ficou conhecido como Carta aos Brasileiros (SILVA, 2002), no qual se disse disposto a discutir com o presidente Fernando Henrique Cardoso (FHC) uma agenda de resposta à crise financeira, o que desagradou setores da esquerda e políticos do próprio PT, partido de oposição à época, fato que pode estar refletido no indicador.

Importante ressaltar que no experimento da Figura 3, o termo oposição foi utilizado conscientemente de forma ampla, sendo classificado como de oposição qualquer partido não pertencente à coalizão de governo, o que é uma simplificação acentuada. Não se trata, portanto, de uma coalizão de oposição, mas de partidos não formadores da base, dos quais não se deve, em princípio, esperar uma convergência ideológica. Apenas para citar um exemplo, esse critério

faz com que, em 2002, o PPS apareça no mesmo pacote do PT. No entanto, sabe-se que naquele ano o candidato à presidência pelo PPS, Ciro Gomes, era, talvez, o mais ferrenho crítico de Lula. De toda forma, essa convergência aparece nos demais anos eleitorais.

Ainda no período FHC, verifica-se forte convergência da base em 2001 e uma situação única de total convergência em 2002. Valer observar que nessa época a coalizão era composta por somente 4 (quatro) partidos, praticamente metade da quantidade que viria a compor as bases dos governos posteriores. Natural aceitar que é mais fácil de se conseguir identidade entre 4 (quatro) do que entre 8 (oito). Quando os diferentes (base e não base) são comparados, nenhum traço de identidade é encontrado.

No período Lula I, os dois primeiros anos mostram queda dos indicadores de base e de oposição, talvez decorrente do fato de que um partido que sempre fora oposição (PT), agora estava no comando e encontraria resistências na própria base (com oito partidos em 2003 e nove partidos em 2004) ao apoiar políticas neoliberais; enquanto que, por outro lado, o partido que comandara o país nos últimos 8 (oito) anos (PSDB), agora assumia, segundo Bezerra (2012), uma oposição programática, de grau oposicionista menor, mas com possibilidade de vitórias em determinadas matérias. Isso também explica o crescimento do indicador dos diferentes, uma vez que, de acordo com a mesma autora, partidos como PSDB e DEM adotaram pontualmente a estratégia de colaborar com o executivo. Em 2005, o indicador de base registra o menor valor no período, em meio a denúncias de corrupção que deflagraram o escândalo do Mensalão, envolvendo partidos da coalizão governista. Nesse mesmo ano, o indicador de oposição cresce, como que decorrente de um resgate de identidade da oposição em torno do discurso de combate à corrupção. Por fim, em 2006, a lógica eleitoral prevalece e tanto base quanto oposição afinam o discurso num movimento crescente de ambos os indicadores, com prevalência para o indicador de base.

O período Lula II mostra certa estabilidade do indicador de base, com pico em 2010, ano eleitoral. Superadas as desconfianças em relação ao primeiro governo Lula, amortecidos os impactos do Mensalão, diante de um cenário de recuperação do crescimento e de economia estável, tudo isso associado à habilidade política do Presidente, parece razoável que não tenha havido grandes variações de identidade no discurso da base. O indicador de oposição, após registrar mínimo em 2007, tem ascendência linear até 2010. O indicador dos diferentes aponta para a similaridade entre os discursos da base e oposição no ano eleitoral, o que condiz com o cenário de alta popularidade de Lula (≈ 87%) e de alto índice de aprovação do governo (≈ 80%). Não seria prudente à oposição pautar o discurso na crítica a Lula, razão pela qual o mote da campanha eleitoral do PSDB foi “O Brasil pode mais”, com as expressões “nós vamos fazer mais” e “podemos fazer mais e melhor” frequentemente repetidas em seus programas eleitorais.

O primeiro ano do período Dilma I repete a tendência natural de queda dos índices de identidade ideológica. Mas a queda do indicador de base não se resumiria a uma tendência

natural. A conjuntura internacional se tornou desfavorável no plano econômico, com baixo crescimento mundial, afetando, também, a economia brasileira. Diante de cenário como esse, a capacidade de articulação política do líder do Executivo é requisito fundamental para se contornar as dificuldades e aprovar as medidas necessárias. Contudo, no campo da habilidade política, a presidente Dilma se mostrou ineficaz, marcada pela incapacidade de dialogar com os atores políticos. Por muitas vezes, deixara de receber deputados e senadores do próprio partido, o que levou o governo a perder votações e espaço para pautar agendas no Congresso Nacional. Em 2012, mesmo sem grandes lançamentos de programas sociais ou números robustos na economia, a popularidade do governo e a avaliação de Dilma bateram recordes, o que, provavelmente, reforçara o perfil autoritário da Presidente. Nesse mesmo ano, o indicador de base atingiu nível inferior ao de 2005, ano do Mensalão. O indicador de oposição também apresentou queda, um pouco menos acentuada. A partir de 2013, os três índices retomam o caminho da lógica das eleições, no qual o indicador de oposição apresenta crescimento expressivo, alcançando valor máximo em 2014, reflexo de uma disputa eleitoral acirrada em que a reeleição de Dilma se deu por diferença de apenas 3,3% dos votos válidos.

No único ano mensurado para o período Dilma II, confirma-se, mais uma vez, a queda dos indicadores. Esse fenômeno decorre, possivelmente, de ajustes naturais motivados pelas novas configurações de mandatos nas esferas municipal, estadual e federal.

7 Considerações finais

O Índice de Identidade Ideológica proposto produziu resultados coerentes com fatos que marcaram a história política brasileira no período compreendido entre 2001 e 2015, sugerindo que o índice possui utilidade analítica para a explicação desses fatos, o que constitui um achado desta pesquisa.

De forma geral, pode-se dizer que o indicador constitui ferramenta objetiva de análise do discurso e que, sob as condições estabelecidas neste estudo, permite investigar o comportamento de alianças políticas, podendo constituir uma forma alternativa à deserção partidária para a aferição da fidelidade. Pode, também, ser usado para a confrontação entre o que o parlamentar fala e como vota, ou seja, se o voto do parlamentar é coerente com o seu discurso.

Nesse sentido, o presente trabalho abre mais um caminho para que o PLN e métodos quantitativos se consolidem como técnicas de análise do discurso parlamentar. Em campo repleto de incertezas, como o da Ciência Política, novos rumos são sempre bem-vindos à elucidação de questões.

Referências

ABRANCHES, S. H. H. DE. Presidencialismo de coalizão: o dilema institucional brasileiro.

Dados - Revista de Ciências Sociais, v. 31, n. 44, p. 5–34, 1988.

AXELROD, R. **Conflict of interest: a theory of divergent goals with applications to politics**. Markham Pub. Co, 1970.

BEZERRA, G. M. L. **A Oposição nos Governos FHC e Lula: um balanço da atuação parlamentar na Câmara dos Deputados**, 2012. Universidade Federal do Rio Grande do Sul. Instituto de Filosofia e Ciências Humanas. Programa de Pós-Graduação em Ciência Política. Disponível em: https://lume.ufrgs.br/handle/10183/70701.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil**. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 20 abr. 2021.

BRASIL. **Emenda Constitucional n. 97 de 4 de outubro de 2017**. Altera a Constituição Federal para vedar as coligações partidárias nas eleições proporcionais, estabelecer normas sobre acesso dos partidos políticos aos recursos do fundo partidário e ao tempo de propaganda gratuito no rádio e na televisão e dispor sobre regras de transição. Brasília, DF: Congresso Nacional, 2017. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc97.htm. Acesso em: 23 abr. 2021.

CÂMARA DOS DEPUTADOS. Centro de Documentação e Informação. **Regimento Interno da Câmara dos Deputados**. 21. ed. Brasília: Edições Câmara, 2020.

CÂMARA DOS DEPUTADOS. Dados Abertos - Legislativo. 2017. Disponível em: https://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo/dados-abertos-legislativo. Acesso em: 1/8/2017.

CANCIAN, R. Ideologia - Termo tem vários significados em ciências sociais. Disponível em: https://educacao.uol.com.br/disciplinas/sociologia/ideologia-termo-tem-varios-significados-em-ciencias-sociais.htm. Acesso em: 13/2/2019.

CEBRAP. **Núcleo de Estudos Comparados e Internacionais – Dados Legislativos**. São Paulo. 2019. Disponível em: http://neci.fflch.usp.br/legislative-data. Acesso em: 25 maio 2018.

CEFOR. **Repositório de Dados Públicos do Programa de Pós-Graduação da Câmara dos Deputados – Discurso Deputados**. Brasília. 2018. Disponível em: https://github.com/Cefor/DiscursoDeputados>. Acesso em: 30 jun. 2021.

EAGLETON, T. **Ideologia: uma introdução**. São Paulo: Editora Boitempo, 1997.

GNERRE, M. **Linguagem, escrita e poder**. 3º ed. São Paulo: Livraria Martins Fontes Editora Ltda., 1991.

GRIMMER, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. **Political Analysis**, v. 18, n. 1, p. 1–35, 2010.

GRIMMER, J.; STEWART, B. M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. **Political Analysis**, v. 21, n. 3, p. 267–297, 2013.

JURAFSKYL, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. New Jersey: Prentice Hall, 2008.

KILGARRIFF, A.; ROSE, T. Measures for corpus similarity and homogeneity. **Proceedings of the Third Conference on Empirical Methods in Natural Language Processing**, p. 46–52, 1998. Disponível em: http://luthuli.cs.uiuc.edu/~daf/courses/Signals AI/Papers/Collocation/kilgarriff98measures.pdf.

MACHADO, C. M.; MIGUEL, L. F. Padrões de coesão e dispersão : Uma proposta de tipologia para coligações. **Teoria & Pesquisa**, v. XX, n. 2, p. 37–58, 2011. Disponível em: https://bibliotecadigital.tse.jus.br/xmlui/handle/bdtse/2962.

MANNING, C. D.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: MIT Press, 1999.

MOREIRA, D. C. **Com a palavra os nobres deputados: frequência e ênfase temática dos discursos dos parlamentares brasileiros**, 2016. Universidade de São Paulo- Faculdade de Filosofia, Letras e Ciências Humanas.

ORLANDI, E. P. **Análise de discurso: princípios e procedimentos**. 12. ed. São Paulo: Pontes Editores, 2015.

QUINN, K. M.; MONROE, B. L.; COLARESI, M.; CRESPIN, M. H.; RADEV, D. R. How to Analyze Political Attention with Minimal Assumptions and Costs. **American Journal of Political Science**, v. 54, n. 1, p. 209–228, 2010.

REISMAN, L. S. **Coalizões, partidos e programas de governo : a submissão das bandeiras partidárias ao mercado eleitoral**, 2016. UNIVERSIDADE DE BRASÍLIA. Disponível em: https://repositorio.unb.br/handle/10482/21469.

RIKER, W. H. **The Theory of Political Coalitions**. Michigan: Yale University Press, 1962.

RODRIGUES, L. M. **Partidos, ideologia e composição social: um estudo das bancadas partidárias na Câmara dos Deputados**. Rio de Janeiro: Centro Edelstein de Pesquisas Sociais, 2009.

ROMA, C. Os efeitos da migração interpartidária na conduta parlamentar. **Dados**, v. 50, n. 2, p. 351–392, 2007.

ROMA, Celso. Os efeitos da migração interpartidária na conduta parlamentar. **Dados**: Revista de Ciências Sociais, Rio de Janeiro, v. 50, n. 2, p. 351-392, 2007. Disponível em: https://doi.org/10.1590/S0011-52582007000200005. Acesso: 09 jun. 2020.

SILVA, E. M. DA; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 19, p. 1–32, 2014. Disponível em: http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/15027/Fundamentos_em_processamento_de_linguagem_natural_uma_proposta_para_extração_de_bigramas.pdf.

SILVA, L. I. L. DA. Leia íntegra da carta de Lula para acalmar o mercado financeiro. **Folha Online**, 22. jul. 2002. Disponível em: https://www1.folha.uol.com.br/folha/brasil/ult96u33908.shtml.

SPIRLING, A. U.S. Treaty Making with American Indians: Institutional Change and Relative Power. **American Journal of Political Science**, v. 56, p. 84–97, 2012.

VYATKINA, N.; BOULTON, A. Corpora in Language Teaching and Learning To cite this version : HAL Id : hal-01237582. **Language Learning and Technology**, v. 21, n. 3, p. 1–8, 2017. Disponível em: https://hal.archives-ouvertes.fr/hal-01237582.

WANG, L.; LIU, R. A Rapid Method to Extract Multiword Expressions with Statistic Measures and Linguistic Rules. **International Conference on Web Information Systems and Mining**, p. 234–241, 2011.

YOUNG, L.; SOROKA, S. Affective News: The Automated Coding of Sentiment in Political Texts. **Political Communication**, v. 29, n. 2, p. 205–231, 2012. Disponível em: https://doi.org/10.1080/10584609.2012.671234.

ZHANGAC, W.; YOSHIDA, T.; TANGB, X.; TU-BAOHOA. Improving effectiveness of mutual information for substantival multiword expression extraction. **Expert Systems with Applications**, v. 36, n. 8, p. 10919–10930, 2009.

Artigo submetido em: 2021-06-11; Artigo aceito em: 2021-06-30