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

Fabiano Peruzzo Schwartz[1]

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 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 Introducción

El trabajo legislativo y la actividad parlamentaria en la Cámara de Diputados se rigen por el Proceso Legislativo[2] y están estrechamente vinculados al diálogo. En el Pleno, máximo órgano de deliberación, los representantes del pueblo debaten y votan soberanamente las propuestas en trámite. La capacidad retórica para ordenar adecuadamente las ideas, proponer los argumentos adecuados, organizar y presentar oralmente el discurso a una audiencia es la principal herramienta del parlamentario en el ejercicio de su oficio. Gnerre (1991, pág. 5) destaca que "el poder de la palabra es el poder de movilizar la autoridad acumulada por el hablante y concentrarla en un acto lingüístico", uno de cuyos ejemplos más elocuentes es el discurso político.

Es de vital importancia comprender la necesidad de un discurso parlamentario y la respectiva influencia en el Proceso Legislativo, que resultará en leyes que afectarán directa o indirectamente la vida de todos los ciudadanos brasileños. Además, entender la dinámica de la política en un amplio espectro solo es posible cuando se sabe qué están diciendo y escribiendo los actores políticos, cómo se posicionan frente a los hechos que afectan a la sociedad. Dado

[1] Doctor en Ingeniería de Sistemas Electrónicos y de Automación y Máster en Ciencias de la Computación, ambos por la Universidad de Brasilia. Director de la Coordinación de Posgrado de la Cámara de Diputados y profesor titular del Máster Profesional en Poder Legislativo. Coordina el grupo de investigación y extensión “Ciencia de datos aplicada al estudio del Poder Legislativo: enfoque computacional y métodos de análisis”, inscrito en el Directorio de Grupos de Investigación del CNPq. ID de Orcid: https://orcid.org/0000-0003-1727-9346. Correo electrónico: fabiano.schwartz@camara.leg.br

[2] Proceso Legislativo es el conjunto de actos realizados por los órganos del Poder Legislativo, siguiendo las reglas establecidas para la elaboración de normas jurídicas tales como reformas de la Constitución, leyes complementarias u ordinarias y otros tipos normativos previstos en el Art. 59 de la Constitución Federal (BRASIL, 1988).

que el discurso es una forma de representar aspectos del mundo en un contexto específico, la articulación de más de uno de ellos es la clave para reunir diferentes perspectivas de comprensión de la realidad. Por lo tanto, el análisis de la acción parlamentaria a través del discurso debe ser considerado por el conjunto del trabajo, no solo por pronunciamientos aislados, ya que estos cargan con el peso del momento y el ambiente en el que se realizaron. Es necesario considerar la secuencia de eventos para identificar la evolución del discurso y de ella extraer conocimiento sobre el proceso político.

Este entendimiento tiene un significado especial en el sistema político brasileño en el que la organización del Poder Ejecutivo se basa en la formación de grandes coaliciones partidistas, lo que fue denominado 'presidencialismo de coalición' (ABRANCHES, 1988). En la práctica, un gobierno necesita formar una base de apoyo en el Congreso Nacional para hacer factibles sus iniciativas de implementación de la política estatal. Por lo tanto, monitorear el comportamiento de esta base en las discusiones de los distintos asuntos discutidos en el parlamento, lo que significa monitorear el comportamiento de los parlamentarios, es una parte intrínseca del juego político. Tanto es así que la Constitución Federal de 1988, en su artículo 17, § 1, destaca que los partidos políticos tienen autonomía para "adoptar los criterios de selección y régimen de sus coaliciones en elecciones de mayoría" y que deben establecer en sus estatutos "normas de disciplina y lealtad al partido" (BRASIL, 1988).

No existe una definición expresa en el texto constitucional de los términos "disciplina" y "fidelidad" partidarias, lo que da lugar a divergencias doctrinales en cuanto a su interpretación y aplicación. La línea de la doctrina que diferencia los términos, vincula: la lealtad partidaria al alineamiento entre parlamentario y partido político en el ámbito ideológico, o el carácter filosófico-programático de los temas; la disciplina partidaria al comportamiento parlamentario frente a las cuestiones del día a día del partido, observada a través de la comparación entre los votos nominales de los parlamentarios y las orientaciones de los líderes del grupo parlamentario. En este sentido, Roma (2007) categoriza dos bloques de parlamentarios: fieles o infieles, cuando el parlamentario, luego de ser elegido, permanece en el partido, o migra a otro; y disciplinados o indisciplinados, cuando las votaciones en el pleno de la Cámara siguen, o no, la dirección de los dirigentes del partido.

Por lo tanto, se espera que las coaliciones estén marcadas por la fidelidad y disciplina partidista de sus miembros. En este aspecto que distingue los términos, la medición de la disciplina tiene un carácter objetivo, bien definido, que permite cuantificar, a través de votos en un período determinado, el porcentaje de veces que el parlamentario o grupo parlamentario sigue la orientación del líder. Por lo tanto, el partido puede monitorear el índice de disciplina de sus parlamentarios a lo largo del tiempo. La fidelidad o, mejor dicho, la infidelidad, se mide en un solo hecho, cuando el parlamentario deserta o abandona el partido, migrando de lista. La identificación de signos de infidelidad a lo largo del tiempo es un proceso subjetivo que debe

considerar otros aspectos del comportamiento parlamentario.

Desde esta perspectiva, el discurso es una materia prima fértil ya que, por definición, tiene un rasgo ideológico. Según Orlandi (2015, pág. 43), las formaciones discursivas, que determinan lo que debe decirse en una coyuntura sociohistórica dada, representan formaciones ideológicas en el discurso, no existiendo en las palabras sentido que no esté determinado ideológicamente. Por tanto, existe una relación recíproca entre ideología y lenguaje.

También existe una relación entre ideología y la lógica de las coaliciones partidistas, en lo que Machado y Miguel (2011) describen como una dimensión programática "ligada a los valores políticos de base, que tiene en cuenta la formación de alianzas entre partidos [...] según su categorización por ideología". Destacan que una coalición es mucho más coherente cuanto mayor es la afinidad de las posiciones ideológicas de sus miembros.

Por tanto, es razonable considerar que del discurso parlamentario se pueden extraer posiciones ideológicas y que la verificación sistemática de estas posiciones es capaz de revelar afinidades ideológicas. Dentro de las coaliciones de partidos también es razonable asociar tales afinidades con un grado de coherencia o identidad. Y por qué no decir cierto grado de fidelidad. En otras palabras, la fidelidad partidista no solo se caracterizaría por el acto de deserción, sino que podría medirse y monitorearse a lo largo del tiempo por lo que los parlamentarios expresan en sus pronunciamientos, que podría ser monitorear al individuo o un grupo de individuos.

El principal problema radica en la dificultad de realizar esta verificación sistemática, ya que requiere un gran esfuerzo para recolectar datos (textos de voz) y analizarlos. Análisis éste que no se restringe al contenido, ya que busca no solo extraer significado del texto discursivo, sino comprenderlo en conjunto con la historia de los hechos. A partir de la recuperación de la línea argumentativa de las actas taquigráficas de los discursos pronunciados en el Pleno de la Cámara de Diputados, la ejecución manual del proceso de verificación resultaría lenta e imprecisa, además de no lograr contemplar la totalidad de los pronunciamientos.

Alternativamente, la captura y gestión del conocimiento implícito en el pronunciamiento parlamentario puede ser mejorado mediante el uso de recursos computacionales y técnicas avanzadas de procesamiento lingüístico. Los científicos políticos han utilizado el análisis automático de contenido en un conjunto diverso de textos. Esto incluye archivos de datos multimedia (YOUNG; SOROKA, 2012); discursos parlamentarios en legislaturas de todo el mundo (QUINN *et al*., 2010); declaraciones del presidente, del legislador y del partido (GRIMMER, 2010); tratados (SPIRLING, 2012); artículos de ciencia política y otros textos políticos.

En este sentido, el presente estudio hace uso de las técnicas de Procesamiento del Lenguaje Natural (PLN) como una herramienta para tratar y ayudar al análisis del discurso parlamentario, y como un recurso útil para establecer conexiones entre hechos, datos cuantitativos y resultados finales del Proceso Legislativo. A partir de la elaboración de datos a

través del PLN, esta investigación innova al proponer un indicador de similitud textual, ahora llamado Índice de Identidad Ideológica, basado en la distribución Qui-Cuadrado (X^2), para mensurar el grado de convergencia en los pronunciamientos de los diputados pertenecientes a un determinado grupo o coalición parlamentarios. Esta medida nos permite investigar, en cierta medida, la solidez de las alianzas políticas o, por extensión, la fidelidad partidista.

El artículo se divide en siete secciones, además de esta Introducción: la segunda sección discute el concepto de ideología político-partidaria y la formación de coaliciones; el tercero presenta técnicas de procesamiento del lenguaje natural; el cuarto, desarrolla las matemáticas para evaluar la similitud del discurso; el quinto propone el Índice de Identidad Ideológica; el sexto analiza los resultados y los principales hallazgos; y el séptimo trae las consideraciones finales.

2 Ideología Político-Partidaria y Coalición

Aquí no se pretende discutir el concepto de "ideología" en su ámbito teórico-filosófico, sino establecer una referencia pragmática, un recurso metodológico con el propósito de identificar hechos, desde el discurso parlamentario, que puedan revelar alguna forma de afinidad, lealtad partidaria y/o coherencia del discurso, capaz de caracterizar una coalición.

La primera teoría sobre el proceso de formación de coaliciones, desarrollada por el politólogo William H. Riker, no presuponía proximidad ideológica (RIKER, 1962). Quizás porque el término ideología es entendido con varios significados, a veces divergentes, esta asociación solo fue propuesta años más tarde por Robert Axelrod (AXELROD, 1970).

Argumentando que una sola definición de ideología sería inútil, Eagleton (1997) presentó una lista de las más comunes, entre las que destacan las siguientes: conjunto de creencias orientadas a la acción; cuerpo de ideas característico de un determinado grupo o clase social; ideas que ayuden a legitimar un poder político dominante; pensamiento de identidad; la coyuntura del discurso y el poder. Según Cancian (2007), en estudios empíricos, el término “ideología” es utilizado con el objetivo de “describir el conjunto de ideas, valores o creencias que orientan la percepción y el comportamiento de los individuos sobre diversos temas o aspectos sociales”, como, por ejemplo, “las opiniones y preferencias que los individuos tienen sobre el sistema político actual, el orden público, el gobierno, las leyes, las condiciones económicas y sociales”. Para ambos autores prevalece en los conceptos el sentido de unidad, de orientación común o identidad, que también se relaciona con el significado de coalición.

Machado y Miguel (2011), al proponer una tipología de coaliciones partidistas, simplificaron el uso del concepto de ideología al asumir que "despojada de sus significados más complejos (y más controvertidos)" la palabra ideología "se refiere únicamente a una posición sobre el eje izquierda-derecha ". Rodrigues (2009, pág. 27), en su estudio sobre grupos partidarios de partidos, discute la lógica ideológica de las coaliciones desde dos ángulos: por un

lado, se cree que las coaliciones reciben una evaluación negativa de la opinión pública porque unen listas ideológica y programáticamente discrepantes, en un escenario de acentuada migración partidista; por otro lado, por el otro, se argumenta que la mayoría de las coaliciones obedecen a la lógica de la afinidad ideológica.

Reisman (2016) destaca que la formación de coaliciones políticas se ubica temporalmente en el momento preelectoral, cuando se negocian alianzas estratégicas entre partidos políticos para elecciones mayoritarias y proporcionales. Al estudiar las coaliciones electorales lideradas por el PT en campañas presidenciales, el autor constató que, si bien los criterios de formación por ideología parecen ser dominantes, esta supuesta convergencia ideológica no impidió el movimiento de las pancartas de las coaliciones desde la izquierda hacia el centro, mostrando que la ideología se somete a la lógica electoral. Tales coaliciones, con el tiempo, sufrieron una drástica reducción de la carga ideológica y adoptaron programas electorales genéricos y abstractos. En su aspecto pragmático, estas coaliciones abandonaron las defensas de puntos controvertidos y se acercaron al centro del eje izquierda-derecha.

El fenómeno observado en el estudio de Reisman (2016), presente en el escenario político general, es probablemente el factor motivador de la Enmienda Constitucional n. 97, del 4 de octubre de 2017, que prohibió las coaliciones de partidos en elecciones proporcionales (BRASIL, 2017), quizás para quitar la lógica electoral del proceso de formación de coaliciones, estimulando la convergencia ideológica en la concepción de los programas.

Por tanto, este estudio propone, de forma simplificada, el uso del concepto “ideología” con el complemento “político-partidaria”, que representa el conjunto de orientaciones políticas con los que se compromete un partido. Aunque este conjunto sufra alteraciones con el tiempo, debe reflejar el compromiso del partido en un contexto temporal determinado. A la luz de la denominación, "ideología político-partidaria", y desde la perspectiva de la formación de coaliciones partidistas, se podría esperar que estas asociaciones se produzcan, en principio, entre partidos que comparten las mismas orientaciones políticas, o al menos de la mayoría de ellas. Asumiendo esta hipótesis como bastante probable, podríamos suponer que dos partidos en coalición pertenecerían, o deberían pertenecer, a la misma ideología político-partidaria.

A partir de estos aspectos, se pueden establecer dos premisas para la realización de este estudio:

1. En principio, una coalición debe basarse en el concepto propuesto de ideología político-partidaria; por tanto, se espera un alto grado de similitud entre los discursos de parlamentarios de partidos en coalición;

2. Un indicador para medir la identidad ideológica de una coalición debe poder captar el comportamiento de todo un grupo de partidos en períodos determinados.

Por tanto, el principal desafío es medir el grado de similitud entre los discursos de los

diferentes parlamentarios de una coalición, lo que requiere un procesamiento de contenido de texto a gran escala, del orden de cientos o miles, una capacidad solo posible a través de recursos informáticos.

Las siguientes dos secciones describen las herramientas técnico-metodológicas utilizadas en este estudio para construir un indicador de identidad ideológica. La técnica de la bolsa de palabras y el método estadístico Qui-Cuadrado son presentados con los siguientes propósitos: preparar las colecciones de textos para su tratamiento computacional; verificar si agrupaciones específicas de dos palabras componen el significado propio de la lengua; y establecer métricas de similitud entre colecciones de texto. Son utilizados el concepto de *corpora*, para hacer referencia a una gran colección de textos digitales sistemáticamente organizados en los que se basa un análisis lingüístico, y el de *corpus* (singular de *corpora*), para un subconjunto de textos extraídos del *corpora* (VYATKINA; BOULTON, 2017).

3 Procesamiento de Lenguaje Natural

El procesamiento de Lenguaje Natural es una subárea de la Inteligencia Artificial y de la Lingüística que estudia los problemas de generación y comprensión automática de lenguas humanas naturales. Grimmer y Stewart (2013), en un estudio de textos políticos, reconocen los beneficios del PLN y cuánto el procesamiento automatizado reduce los costos y esfuerzos para analizar grandes colecciones de texto. Sin embargo, enfatizan que los métodos automatizados no sustituyen un pensamiento cuidadoso y una lectura cuidadosos, además de exigir una validación extensa y específica del problema. Argumentan que, para que el procesamiento automatizado de texto se convierta en una herramienta estándar para los científicos políticos, los investigadores en el campo deben contribuir a crear formas sólidas de validación de los métodos.

Zhang *et al* (2009) afirman que la capacidad de expresar el sentido de una palabra depende de las demás palabras que la acompañan. Cuando una palabra aparece acompañada de un conjunto de términos, mayores son las posibilidades de que ese conjunto tenga un significado relevante. Esto significa que no solo la palabra, sino también la información contextual es útil para el procesamiento de la información.

Silva y Souza (2014) enfatizan que el texto no es un simple conjunto de palabras al azar, sino que el orden en el que son reunidas produce el significado. El estudio de la coocurrencia de palabras puede indicar si están relacionadas directamente, por composición o afinidad, o indirectamente, por semejanza. Por tanto, la base de la lingüística empírica es encontrar, a partir de la frecuencia de coocurrencias observadas, las dependencias significativas entre los términos. Estos términos adyacentes son denominados n-gramas, donde *n* es la cantidad de términos.

Los modelos de lenguaje n-gram fueron desarrollados en el campo de la Estadística por el matemático ruso Andrey Markov (1856-1922) con el fin de reconocer patrones estadísticos de

uso de la lengua basados en cadenas, conocidas como cadenas de Markov. Wang y Liu (2011) señalan que muchos trabajos tienen como foco dominante la identificación y extracción de n-gramas. Este proceso es denominado tokenización o segmentación de palabras y consiste en la tarea de dividir un texto en unidades mínimas llamadas tokens, donde cada token puede ser una palabra o conjunto de palabras, un número, un símbolo de puntuación u otra estructura perteneciente al lenguaje (MANNING; SCHÜTZE, 1999).

La complejidad de la tokenización varía según la complejidad del propio idioma, con la definición del problema a ser estudiado y con el modelo elegido para aplicación del PLN. El proceso de tokenización generalmente incluye las siguientes etapas: conversión de los caracteres de texto a minúsculas; eliminación de números, puntuación, plurales, acentos, *stopwords*[3] y espacios en blanco; y segmentación de las palabras. El producto final de la tokenización es la bolsa de palabras (en inglés, *bag of words*) con el recuento de frecuencia de los términos. En el enfoque de la bolsa de palabras, no se considera el orden en que las palabras están dispuestas en el texto y la búsqueda de información solo tiene en cuenta las frecuencias estimadas. La Figura 1 ilustra un fragmento de discurso antes y después del proceso de tokenización. La significativa reducción del texto se debe principalmente al conjunto de *stopwords* utilizado.

[3] *Stopwords* son palabras con poco significado en algunas aplicaciones de PLN, como recuperación y clasificación de información, lo que significa que estas palabras no son muy discriminatorias. Los artículos y pronombres generalmente son clasificados como *stopwords*. La idea es simplemente eliminar las palabras que aparecen con mucha frecuencia en todos los documentos de la colección de texto. Sin embargo, la elección de la lista de *stopwords* debe ser cuidadosa, ya que es determinante para lograr los objetivos.

Figura 1 – Tokenización aplicada a trecho del discurso parlamentario: bolsa de palabras.

Extracto Original (versión em Español):

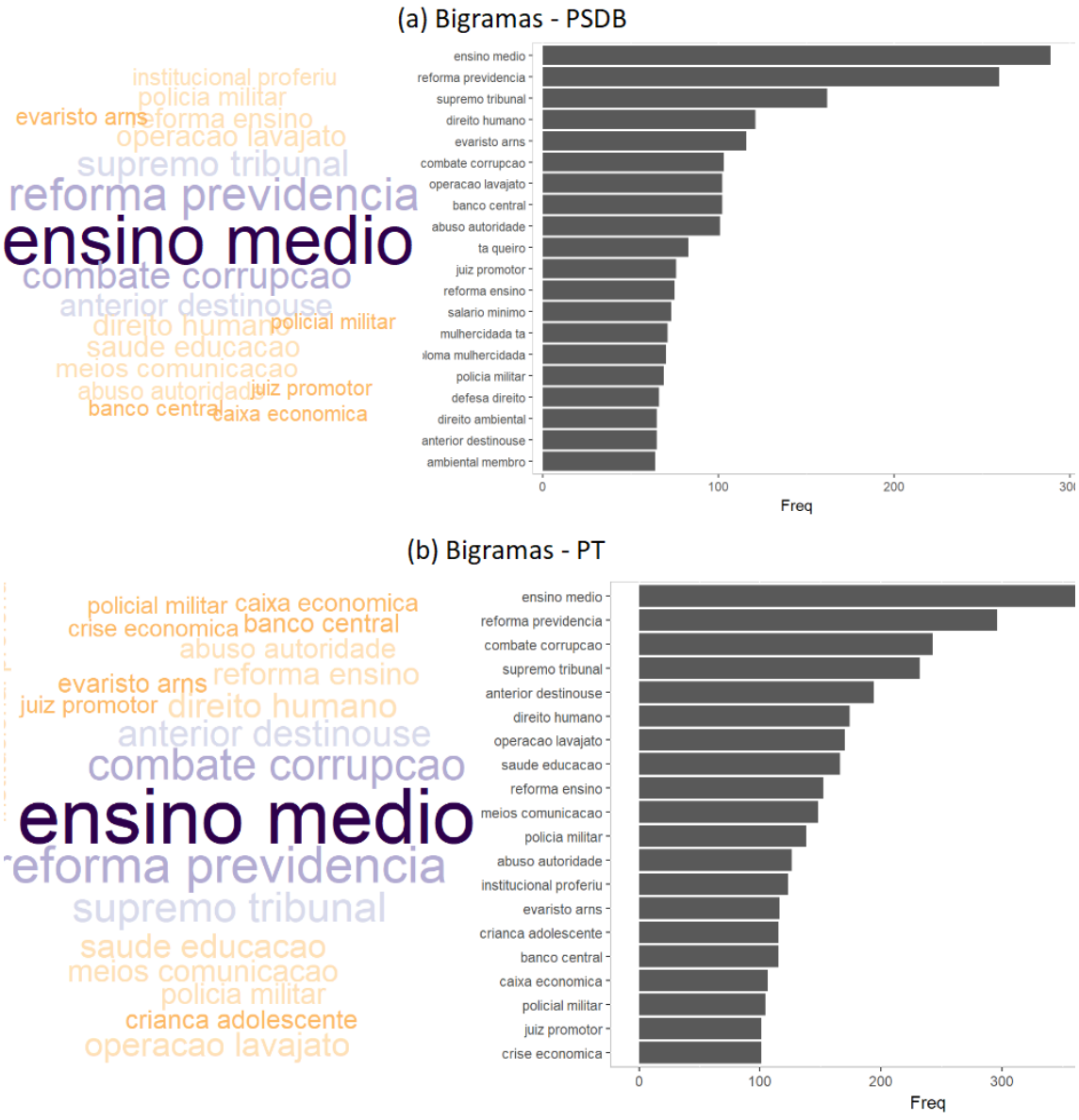
Presentamos enmiendas para hacer supresión. Entonces, sí, tendríamos los recursos necesarios en reserva para el aumento del salario mínimo. Hicimos un acuerdo con el Líder de Gobierno, Diputado Fulano de Tal, en el que tendríamos una reserva de 6 mil millones de reales, con mil millones de reales para la Bolsa-Família.

Fuente: Elaboración del autor.

El proceso de tokenización puede ser ajustado para efectuar la división del texto en bigramas, es decir, tokens con dos palabras que aparecen consecutivamente en el texto. Los bigramas son comúnmente utilizados como base para el análisis estadístico de textos y ayudan a estimar la probabilidad condicional de una palabra, dada la palabra precedente (JURAFSKY; MARTIN, 2008). Los bigramas son más ricos en significado que el conteo de palabras simples y son más útiles para extraer información del texto.

La Figura 2 ilustra el conjunto de bigramas extraídos de los discursos pronunciados por los diputados del PSDB y del PT en el año 2016, presentados en la forma de nubes de palabras y gráficos de frecuencia. Estos recursos constituyen un método heurístico de análisis que, por sí solo, no permite generalizar la respuesta a una cuestión de investigación, pero señala formas de lo que debe o no recibir más atención en un conjunto de textos.

Figura 2 – Nube de palabras y gráfico de barras de frecuencias absolutas: (a) bigramas de los discursos del PSDB en 2016; (b) bigramas de los discursos del PT en 2016.



Glosario:

abuso autoridade	abuso de autoridad
banco central	banco central brasileiro
caixa economica	caja de ahorros federal
combate corrupcao	combatir la corrupción
crianca adolescente	niño y adolescente
crise economica	crisis económica
defesa direitos	defensa de derechos
direito ambiental	derecho ambiental
direito humano	derecho humano
ensino medio	escuela secundaria
evaristo arns	Evaristo Arns fue un fraile franciscano brasileño
juiz promotor	fiscal de distrito
meios comunicacao	medios de comunicación
opercao lavajato	La Operación Lava Jato fue una investigación criminal de la Policía Federal de Brasil
policia militar	policia militar
reforma ensino	reforma educativa
reforma previdencia	reforma de la seguridad social
salario minimo	salario mínimo
saude educacao	salud y educacion
supremo tribunal	Corte Suprema

Fuente: Elaboración del autor.

En una primera mirada de la Figura 2, contrariamente a lo que el sentido común podría esperar de los pronunciamientos de partidos abiertamente opuestos, uno tiene la impresión de que los discursos son similares y abordan los mismos temas, con ligeras variaciones. Sin embargo, aunque importante para la maduración del proceso investigativo, la inspección visual no es suficiente para llegar a conclusiones, requiriendo enfoques más robustos a través de pruebas estadísticas, que serán detallados en la siguiente sección.

4 Similitud del discurso por comparación de bigramas como *collocations*

La tarea de mensurar cuánto un *corpus* de discursos se asemeja a otro mediante un método computacional debe preservar, en la medida de lo posible, los significados de los textos. A pesar de que la técnica de la bolsa de palabras no tiene la función de evaluación semántica, ya que ignora el orden en que las palabras están dispuestas en el texto, la división en n-gramas con *n* mayor que 1 (n > 1) da como resultado conjuntos de términos cuyo orden coincide con el del texto original. Por tanto, estos n-gramas llevan más información semántica que términos aislados, y es razonable suponer, por ejemplo, que la comparación entre discursos basados en bigramas es más efectiva que la basada en términos aislados.

Este supuesto cobra aún más fuerza si somos capaces de identificar bigramas cuyos términos no ocurren juntos por casualidad, sino por alguna razón de dependencia que exprese una forma convencional, en el lenguaje, de decir las cosas, esto es, que se inserte en un contexto semántico. Las expresiones de dos o más palabras que corresponden a una forma convencional de decir las cosas se denominan colocaciones, o *collocations* (MANNING; SCHÜTZE, 1999).

La prueba estadística de qui-cuadrado (X^2) puede ser utilizada para evaluar si dos palabras constituyen una *collocation*, verificando la dependencia entre ellos con la ayuda de una tabla de contingencia (KILGARRIFF; ROSE, 1998). En esencia, la prueba trata cada palabra como una variable categórica y asume la condición de independencia de los eventos, es decir, cuando las palabras suceden juntas al azar y, por tanto, no constituyen una *collocation*. Por definición, dos eventos son independientes cuando la probabilidad de que ambos ocurran juntos es igual al producto de las probabilidades de que cada evento ocurra individualmente:

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

Entonces, el valor esperado para la ocurrencia conjunta de dos eventos independientes viene dado por

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

donde N es el número total de eventos.

Por lo tanto, conociendo el valor esperado de eventos independientes, la estadística X^2 es determinada comparándose las frecuencias observadas con las frecuencias esperadas. Cuando

no existe diferencia estadística entre las frecuencias observadas y las esperadas, se concluye que los eventos son independientes. Las hipótesis bajo la condición de independencia de los términos del bigrama son enunciadas por:

H_0: No hay asociación entre los términos; el bigrama no es una *collocation*.

H_1: Existe una asociación entre los términos; el bigrama es una *collocation*.

En la práctica, son computadas la frecuencia con la que las palabras del bigrama ocurren juntas, las frecuencias con las que forman bigramas con otras palabras y la frecuencia de los bigramas formados sin estas palabras. La Tabla 1 muestra la distribución de los términos "enseñanza" y "media" (sin acento después de la tokenización) para el análisis del bigrama "enseñanza media", donde w_1 y w_2 se refieren, respectivamente, a la primera y segunda palabra del bigrama.

Tabla 1 – Tabla de contingencia: distribución de los términos “ensino” y “medio” para el análisis del 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: frecuencias medidas de los discursos de PT en 2016; “ensino medio” significa “escuela secundaria”; “oriente medio” significa “Medio Oriente”; “ensino superior” significa "educación universitaria".

Fuente: Elaboración propia.

Luego, a partir de la tabla de contingencia, se estima la estadística X^2 de la siguiente manera:

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

Donde *i* representa las líneas y *j* las columnas de la tabla, O_{ij} es el valor observado en la celda (*i*, *j*) y E_{ij} es el valor esperado. Os valores de E_{ij} son computados conforme a Ecuación 2 a partir de las probabilidades marginales, esto es, a partir de los totales de líneas y columnas convertidos en proporciones. Por ejemplo, para la celda (1, 1), el valor esperado E_{11} viene dado por el producto de la probabilidad de que "enseñanza" sea el primer término del bigrama ($P(A)$), la probabilidad de que "medio" sea el segundo término del bigrama ($P(B)$) y la cantidad total de bigramas (N) existentes en el *corpus* en análisis:

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

Aplicando la Ecuación 3 a todas las celdas de la Tabla 1, encontramos $X^2 \approx 28{,}75$ y $p \approx 0.0000000821$, lo que significa que H_0 debe ser rechazada, es decir, las palabras "enseñanza" y "media" no ocurren juntas al azar. Son, por ende, dependientes y constituyen una *collocation*.

Una vez identificadas las *collocations* de dos *corpus*, la similitud entre ellos puede ser medida, una vez más, a través de estadísticas X^2. Kilgarriff y Rose (1998) propusieron este enfoque en un estudio donde compilaron una tabla de *n* líneas por 2 columnas (con $n = 500$), en la cual cada columna correspondía al recuento de las palabras más frecuentes comunes a los dos *corpus*. En el presente estudio, el método de Kilgarriff y Rose (1998) fue extendido a bigramas validados como *collocations*. La Tabla 2 contiene una representación esquemática de esta concepción, utilizando el *corpus* de los discursos del PT y PSDB en 2016.

Tabla 2 – *Collocations* más frecuentes comunes a los *corpus* de discursos del PT y PSDB en 2016.

	collocation	**frec. PT**	**frec. 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	....	....	....

Nota: "ensino medio" significa "escuela secundaria"; "reforma previdencia" significa "reforma de la seguridad social"; "combate corrupcao" significa "combatir la corrupción"; "supremo tribunal" significa "Corte Suprema"; "direito humano" significa "derecho humano"; "opercao lavajato" fue una investigación criminal de la Policía Federal de Brasil; "saude educacao" significa "salud y educacion"; "reforma ensino" significa "reforma educativa"; "meios comunicacao" significa "medios de comunicación"; "policia militar" significa "policia militar".

Fuente: Elaboración propia.

Para cada uno de los *n collocations* más frecuentes en la Tabla 2 es necesario calcular el número de ocurrencias esperadas (valor esperado) en cada *corpus*, con el fin de verificar si ambos pueden ser considerados muestras aleatorias de una misma población. Dados los tamaños N_1 y N_2, de los *corpus* 1 y 2, y las frecuencias observadas O_{i1} y O_{i2} de la *i*-ésima *collocation* en los respectivos *corpus*, entonces, el valor esperado para esta *collocation* en el *corpus* 1 viene dado por la probabilidad de que ocurra una *collocation* en el conjunto total *corpus* 1 más *corpus* 2) multiplicado por el tamaño del *corpus* 1 (N_1)

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

y respectivamente en el *corpus* 2, por

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

Conociendo los valores observados y esperados, entonces X^2 puede ser determinado por la Ecuación 7,

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

Donde *C* representa el respectivo *corpus*. Las hipótesis nula y alternativa pueden ser formuladas de la siguiente manera:

> H_0: Las frecuencias observadas no difieren de las frecuencias esperadas, es decir, no existe diferencia entre las frecuencias de los grupos.
>
> H_1: Las frecuencias observadas son diferentes de las frecuencias esperadas, por lo tanto, existe una diferencia entre las frecuencias de los grupos.

Por lo tanto, aceptar H_0 significa decir que no existe diferencia entre los *corpus*.

Una vez definidas las herramientas computacionales y estadísticas, la siguiente sección aplica los conceptos explorados en la construcción del indicador de similitud propuesto, con base en el concepto de ideología político-partidista.

5 Índice de identidad ideológica

A partir de la propuesta de Kilgarriff y Rose (1998) extendida a los bigramas, se realizó una primera prueba de similitud para combinaciones de los partidos. PT, PSDB, PMDB[4], PSOL, PCDOB y PTB, cuyos *corpus* del año 2016 fueron comparados de dos en dos a través de las respectivas *collocations* (Tabla 3).

[4]En el período correspondiente a la estructuración del *corpora*, el partido todavía usaba el acrónimo PMDB, que luego cambió a MDB.

Tabla 3 – Métrica X^2 para similitud de corpus: partidos PT, PSDB, PMDB, PSOL, PCdoB y PTB; año 2016.

Partido 1	Partido 2	Bigrama (*collocation*)		
		X^2	*p*	H_0
PT	PSDB	2.386,57	7,24e-302	rechaza
PT	PMDB	1.836,77	1.67e-196	rechaza
PT	PSOL	4.093,27	0,00	rechaza
PT	PCdoB	4.217,40	0,00	rechaza
PT	PTB	4.390,26	0,00	rechaza
PSDB	PMDB	165,68	1,00	acepta
PSDB	PSOL	1.679,38	1,91e-221	rechaza
PSDB	PCdoB	1.763,91	2,38e-238	rechaza
PSDB	PTB	3.289,55	0,00	rechaza
PMDB	PSOL	1.919,91	7,75e-277	rechaza
PMDB	PCdoB	1.994,99	1,83e-292	rechaza
PMDB	PTB	3.113,44	0,00	rechaza
PSOL	PCdoB	16,83	1,00	acepta
PSOL	PTB	708,59	8,29e-32	rechaza
PCdoB	PTB	648,03	8,91e-23	rechaza

Fuente: Elaboración del autor.

Se observa en la Tabla 3 la similitud de *corpus* en combinaciones PSDB/PMDB y PSOL/PCdoB. En un análisis preliminar, esto parece razonable en un año en el que *el proceso de destitución* de la presidenta Dilma dominó el escenario político, en el que partidos como el PSDB y el PMDB se aliaron con la causa, mientras que otros, como el PSOL y el PCdoB, tomaron la postura contraria. También se observa que el discurso del PT no es similar al de ningún otro partido, reflejando quizás una ruptura en la base de gobierno ante los numerosos intereses que rodearon el tema *proceso de destitución*.

Este primer experimento demostró ser coherente en la comparación de los *corpus*, revelando la capacidad explicativa de los acontecimientos, que inspiró la construcción de un indicador con la función de mensurar el grado de similitud en el discurso de un grupo de partidos afines, denominado **Índice de identidad ideológica**. La idea es **determinar la probabilidad media de las estadísticas X^2 encontrado en la comparación de los discursos de partidos pertenecientes a un grupo dado, tomados de dos en dos**. Debido al presupuesto asumido de la ideología partido-político, se debe esperar que este indicador tenga un valor muy cercano a 1 (uno) cuando mensurado en partidos de la misma coalición. Por analogía, cuando se

aplica a partidos con orientaciones políticas opuestas, debe producir valores cercanos a cero.

A los efectos de este estudio, fueron considerados los discursos pronunciados en el Pequeño[5] y Grande expediente[6]. La elección de estas fases de la sesión ordinaria del Pleno se debe a que favorecen un debate ideológico, menos influido por la presión de las votaciones y de los ánimos políticos caldeados.

Al respecto, Moreira (2016) identificó que los discursos pronunciados en el Pequeño Expediente no presentan una concentración de temas, “hay indicios de que esta estrategia está especialmente orientada por la cantidad de discursos pronunciados y por la ideología de la lista”. El autor agrega que "existen diferencias entre los parlamentarios en el enfoque temático" y que "esta diferencia está influenciada por el porcentaje de votos recibidos, por la ideología de la lista del partido por el que el parlamentario fue electo".

Este fenómeno puede deberse a la libertad que tiene el parlamentario durante el Pequeño Expediente, pudiendo hablar sobre temas libres por hasta 5 (cinco) minutos. También según Moreira (2016), "muchos diputados pasan todo el mandato sin mencionar votaciones y comisiones, pero en el Pequeño Expediente casi todos hablan", mostrando que, en las legislaturas de número 51 a 54, el 88% de los diputados hizo al menos un discurso en el Pequeño Expediente, con una media de 59 discursos por diputado. Asimismo, se incluyó el Grande Expediente porque tiene dos similitudes con el Pequeño Expediente: el diputado expresa su intención de hablar mediante inscripción; hay libertad para el tema del discurso.

6 Resultados y discusión

Se realizaron tres mediciones distintas del Índice de Identidad Ideológica, año a año, en el período de 2001 a 2015: la primera considera solo a los partidos pertenecientes a la coalición de gobierno (o base gobernante); la segunda considera únicamente a los partidos que no pertenecen a la coalición de gobierno (aquí definidos, en general, como oposición); la tercera, estima la similitud del discurso entre partidos pertenecientes y no pertenecientes a la coalición de gobierno. Los datos sobre la composición de las coaliciones de gobierno cada año fueron extraídos de la Base de Datos Legislativa del Centro Brasileño de Análisis y Planificación - Cebrap (CEBRAP, 2019). Todo el contenido de los discursos pronunciados en el Pleno de la Cámara, en ese período, fue descargado del Portal de Datos Abiertos de la Cámara de Diputados (CÁMARA DE DIPUTADOS, 2017). En total, se leyeron y procesaron más de 150 mil discursos, según datos y códigos de programación en Lenguaje R disponibles en un repositorio público (CEFOR, 2018).

[5] Primera parte de la sesión ordinaria del Pleno, tiene una duración máxima de 60 minutos y está destinada a comunicaciones de parlamentarios previamente inscritos (CÁMARA DE DIPUTADOS, 2020, pág. 48).

[6] Fase de la sesión plenaria que sigue a la del Pequeño Expediente, con una duración improrrogable de cincuenta minutos. Está destinado al discurso de los parlamentarios inscritos por hasta veinticinco minutos por cada orador, incluyendo los eventuales apartados concedidos (CÁMARA DE LOS DIPUTADOS, 2020, pág. 49-50).

Para cada año del período, se generó una matriz de similitud como se muestra en la Tabla 4, que muestra los valores estimados para el año 2003. Los partidos escritos en mayúsculas corresponden a la coalición gobernante.

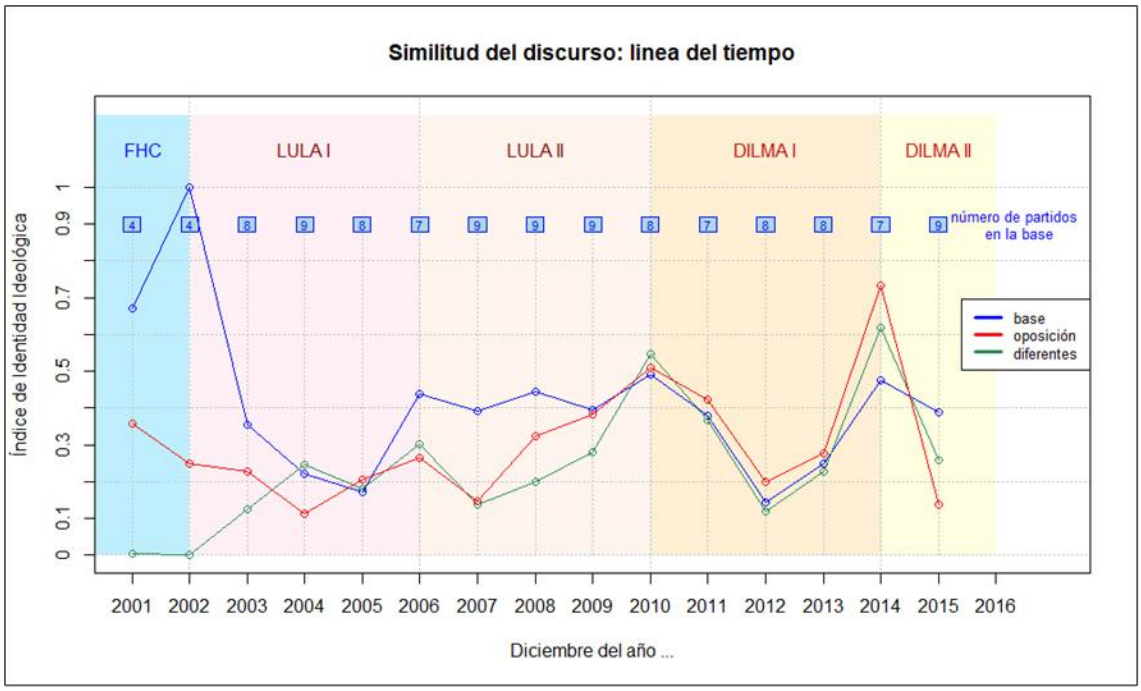
Tabla 4 – Matriz de similitud discursiva del año 2003: $\mathbf{I_{coa}}$ es el índice de identidad ideológica de la coalición gobernante; $\mathbf{I_{opo}}$ es el índice de identidad ideológica de la oposición; $\mathbf{I_{dif}}$ es el índice de identidad ideológica cuando se comparan diferentes grupos.

	pfl	**PCdoB**	**PDT**	**PMDB**	**PP**	**PPS**	**PR**	**prona**	**PSB**	**PSC**	**PSDB**	**PT**	**PTB**	**PV**	$\mathbf{I_{coa}}$	$\mathbf{I_{opo}}$	$\mathbf{I_{dif}}$
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	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	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
														Promedio	**0,36**	**0,23**	**0,12**

Fuente: Elaboración del autor.
Nota: los partidos escritos en mayúsculas corresponden a la coalición gobernante; las celdas entre partidos representan la probabilidad de la estadística X^2 resultante de la comparación de los respectivos partidos; se destacan dos líneas como ejemplo, una para la coalición (PCdoB) y otra para la oposición (pmdb); las celdas en azul destacan los términos usados en la línea respectiva para estimar el $\mathbf{I_{coa}}$; las celdas en salmón destacan los términos usados en la línea respectiva para estimar el $\mathbf{I_{opo}}$; las celdas en color verde destacan los términos usados en la línea respectiva para estimar el $\mathbf{i_{dif}}$.

La Figura 3 ilustra los resultados encontrados para el indicador en el período 2001 a 2015, destacando los respectivos gobiernos en ese momento.

Figura 3 – Índice de identidad ideológica en el período de 2001 a 2015.



Fuente: Elaboración del autor.

Es posible identificar una relación coherente entre los resultados de la Figura 3 y los hechos ocurridos en cada fase de gobierno, con el fin de evaluar el potencial analítico del indicador propuesto. Un primer patrón observado es el aumento del indicador base gobernante en los años electorales (2002, 2006, 2010 y 2014), seguido de una caída en el año siguiente a la elección. Este comportamiento tiende a confirmar el argumento de Reisman (2016) de que la ideología se somete a la lógica electoral. Un movimiento similar se observa en el indicador de oposición, excepto para 2002, año en el que Lula sería electo para su primer mandato como Presidente de la República. En esa ocasión, cuando Lula empezó a crecer en las encuestas de opinión, se extendió un clima de inseguridad en el ámbito político y económico, bajo la sospecha de que el país pudiera ir a la quiebra. Lula se vio obligado a firmar un texto que se conoció como la Carta a los Brasileños (SILVA, 2002), en el que dijo estar dispuesto a discutir con el presidente Fernando Henrique Cardoso (FHC) una agenda para responder a la crisis financiera, lo que desagradó a sectores de izquierda y políticos del propio PT, partido de oposición en ese momento, hecho que puede verse reflejado en el indicador.

Es importante resaltar que en el experimento de la Figura 3, el término oposición fie utilizado conscientemente de manera amplia, siendo clasificado como oposición cualquier partido no perteneciente a la coalición de gobierno, lo cual es una simplificación acentuada. Por tanto, no se trata de una coalición de oposición, sino de partidos que no forman la base, de los que, en principio, no se debe esperar una convergencia ideológica. Por citar sólo un ejemplo, este criterio hace que, en 2002, el PPS aparezca en el mismo paquete que el PT. Sin embargo, se

sabe que, en ese año, el candidato presidencial por el PPS, Ciro Gomes, fue quizás el crítico más ferviente de Lula. De todos modos, esta convergencia aparece en otros años electorales.

Aún en el período de FHC, hubo una fuerte convergencia de la base en 2001 y una situación única de convergencia total en 2002. Cabe señalar que en ese momento la coalición estaba compuesta por solo 4 (cuatro) partidos, prácticamente la mitad de lo que compondría las bases de los gobiernos posteriores. Es natural aceptar que es más fácil obtener identidad entre 4 (cuatro) que entre 8 (ocho). Cuando se comparan los diferentes (base y no base), no se encuentra ningún rasgo de identidad.

En el período Lula I, los dos primeros años muestran una caída en los indicadores de base y de oposición, quizás por el hecho de que un partido que siempre había sido la oposición (PT) ahora estaba al mando y encontraría resistencias en la propia base (con ocho partidos en 2003 y nueve partidos en 2004) en apoyar políticas neoliberales; mientras que, por otro lado, el partido que había comandado el país durante los últimos 8 (ocho) años (PSDB), ahora asumía, según Bezerra (2012), una oposición programática, de menor grado de oposición, pero con la posibilidad de victorias en determinadas materias. Esto también explica el crecimiento del indicador de los diferentes, ya que, según la misma autora, partidos como el PSDB y el DEM adoptaron puntualmente la estrategia de colaborar con el ejecutivo. En 2005, el indicador base registra el valor más bajo del período, en medio de acusaciones de corrupción que desencadenaron el escándalo Mensalão, involucrando a partidos de la coalición gobernante. En ese mismo año, el indicador de oposición creció, como por un rescate de la identidad de la oposición en torno al discurso anticorrupción. Finalmente, en 2006 prevalece la lógica electoral y tanto la base como la oposición agudizan el discurso en un movimiento creciente de ambos indicadores, con predominio del indicador de base.

El período Lula II muestra cierta estabilidad en el indicador de base, con un pico en 2010, año electoral. Una vez superadas las desconfianzas sobre el primer gobierno de Lula, amortiguados los impactos del Mensalão, ante un escenario de recuperación del crecimiento y una economía estable, todo esto asociado a la habilidad política del presidente, parece razonable que no hubiera grandes variaciones en la identidad en el discurso de la base. El indicador de oposición, tras registrar un mínimo en 2007, tiene un ascenso lineal hasta 2010. El indicador de los diferentes apunta a la similitud entre los discursos de base y de oposición en el año electoral, lo cual se condice con el escenario de alta popularidad de Lula (≈ 87%) y un alto índice de aprobación del gobierno (≈ 80%). No sería prudente que la oposición basara su discurso en críticas a Lula, razón por la cual el lema de la campaña electoral del PSDB fue “Brasil puede hacer más”, con las frecuentes expresiones “vamos a hacer más” y “podemos hacer más y mejor” frecuentemente repetidas en sus programas electorales.

El primer año del período Dilma I repite la tendencia descendente natural de los índices de identidad ideológica. Pero la caída del indicador de base no se reduciría a una tendencia

natural. La coyuntura internacional se tornó desfavorable en el plano económico, con un bajo crecimiento global, afectando también a la economía brasileña. En un escenario como este, la capacidad de articulación política del líder del Ejecutivo es un requisito fundamental para superar las dificultades y aprobar las medidas necesarias. Sin embargo, en el campo de la habilidad política, la presidenta Dilma resultó ineficaz, marcada por su incapacidad para dialogar con los actores políticos. En muchas ocasiones no logró recibir diputados y senadores del propio partido, lo que llevó al gobierno a perder votos y espacio para fijar agendas en el Congreso Nacional. En 2012, incluso sin grandes lanzamientos de programas sociales o cifras sólidas en la economía, la popularidad del gobierno y la evaluación de Dilma batieron récords, lo que probablemente reforzó el perfil autoritario de la presidenta. En ese mismo año, el indicador base alcanzó un nivel por debajo del de 2005, año del Mensalão. El indicador de oposición también presentó una caída, un poco menos acentuada. A partir de 2013, los tres índices retornan a la lógica de las elecciones, en las que el indicador de oposición muestra un crecimiento significativo, alcanzando un valor máximo en 2014, reflejo de una feroz disputa electoral en la que la reelección de Dilma se debió a una diferencia de solo 3.3 % de los votos válidos.

En el único año medido para el período Dilma II, una vez más se confirma la caída de los indicadores. Este fenómeno posiblemente sea el resultado de ajustes naturales motivados por nuevas configuraciones de mandatos a nivel municipal, estadual y federal.

7 Consideraciones finales

El Índice de Identidad Ideológica propuesto arrojó resultados consistentes con hechos que marcaron la historia política brasileña en el período entre 2001 y 2015, sugiriendo que el índice tiene utilidad analítica para explicar estos hechos, lo que constituye un hallazgo de esta investigación.

En general, se puede decir que el indicador es una herramienta objetiva para el análisis del discurso y que, en las condiciones establecidas en este estudio, permite investigar el comportamiento de las alianzas políticas, que pueden constituir una forma alternativa a la deserción partidista para medir la fidelidad. También puede ser utilizado para el enfrentamiento entre lo que dice el parlamentario y cómo vota, es decir, si el voto del parlamentario es coherente con su discurso.

En este sentido, el presente trabajo abre otra vía para que el PLN y los métodos cuantitativos se consoliden como técnicas de análisis del discurso parlamentario. En un campo lleno de incertidumbres, como las Ciencias Políticas, nuevos rumbos siempre son bienvenidos para elucidar cuestiones.

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.