PARLIAMENTARY SPEECH SIMILARITY INDICATOR: ANALYSIS OF BEHAVIOR OF PARTY COALITIONS

Fabiano Peruzzo Schwartz[1]

Abstract: 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.

1 Introduction

Legislative works and the parliamentary activity in the Chamber of Deputies are governed by the Legislative Process[2] and intricately connected to dialog. In the Plenary, the highest decision-making body, the peoples' representatives sovereignly discuss and vote on the proposals in process. The rhetorical ability to appropriately arrange ideas, place adequate arguments, organize and orally present the discourse for an audience is the main parliamentary tool in the exercise of this activity. Gnerre (1991, p.5) highlights that "the power of the word is the power to mobilize the authority accumulated by the speaker and concentrate it in a linguistic act," one of the most eloquent examples is political discourse.

Realizing the need for parliamentary speech and the respective influence in the Legislative Process, from which will result laws that will directly or indirectly affect the lives of all Brazilian citizens, is of utmost importance. Furthermore, understanding the dynamics of politics in a broad spectrum is only possible when one has knowledge of what political actors are saying and writing, how they position themselves in the face of facts that affect society. Since speech is a way of representing world aspects in a specific context, articulation of more

[1] PhD in Engineering and Electronic and Automation Systems, Master in Computer Science, both by Universidade de Brasília. Director of the Chamber of Deputies' Graduate program coordination and full professor of the Professional Master's in Legislative Power. Coordinates the research and extension group "Data Science Applied to the Study of Legislative Power: computational approach and analysis methods," registered in the CNPq Research Groups Directory Orcid Id: https://orcid.org/0000-0003-1727-9346. Email: fabiano.schwartz@camara.leg.br

[2] Legislative Process is the set of acts performed by the Legislative Branch's bodies, following fixed rules to develop legal rules such as Constitutional amendments, complementary or ordinary laws and other normative types provided for in art. 59 of the Federal Constitution (BRASIL, 1988).

than one of them is key to bringing together different perspectives to understand reality. Therefore, analysis of parliamentary action through speech should be considered for its whole, not just by isolated pronouncements, for these carry the weight of the moment and the environment in which they were made. It is necessary to consider the sequence of events to identify the evolution of the speech and extract knowledge about the political process from speech.

This understanding has a special meaning in the Brazilian political system where the organization of the Executive Branch is based on the formation of large party coalitions, which were named "coalitional presidentialism" (ABRANCHES, 1988). In practice, a government needs to form a support base in Congress to make its initiatives to implement state policy viable. However, monitoring the behavior of this base in discussions on the several issues argued in parliament, which means monitoring the behavior of parliamentarians that comprise it, is the intrinsic part of the political game. So much so, that the Federal Constitution of 1988, in its article 17, paragraph 1, emphasizes that political parties have the autonomy to "adopt the selection criteria and the regime of their coalitions in majority elections" and that they must establish "standards of discipline and party loyalty" in their bylaws (BRASIL, 1988).

There is no express definition in the constitutional text of the terms party "discipline" and "loyalty," which leads to doctrinal divergences regarding their interpretation and application. The doctrinal line that differentiates the terms links: party loyalty to the alignment between parliamentarians and the political party in the ideological sphere, or philosophical-programmatic character of the themes, party discipline to the parliamentarian's behavior in face of the party's day-to-day issues, observed through the comparison between the roll-call votes of parliamentarians and the guidelines of the party leaders. In this sense, Roma (2007) categorizes two blocks of parliamentarians: faithful or unfaithful, when the parliamentarian, after being elected, remains in the party, or migrates to another; and disciplined or undisciplined, when the votes in the plenary of the Chamber follow, or not, the guidance of the party leaders.

Therefore, expectation is that coalitions be marked by their members' partisan loyalty and discipline. In this line of thought that distinguishes the terms, measuring discipline has an objective, well defined character, that allows quantification of the percentage of times that the parliamentarian or their caucus has followed the party leader's guidance, using votes in a given period. Thus, the party can monitor the discipline rate of its parliamentarians over time. Loyalty, or disloyalty, is measured in a single fact, when the parliamentarian deserts or abandons the party, migrating to another. Identifying signs of disloyalty over time is a subjective process, that should observe other aspects of parliamentary behavior.

From this perspective, speech is a fertile input as, by definition, it carries ideological traits. According to Orlandi (2015, p. 43), formations in discourse, which determine what should be said in each socio-historical context, represent ideological formations in the

discourse, with no meaning in words that is not ideologically. Thus, there is a reciprocal relationship between ideology and language.

There is also a relationship between ideology and the logic of party coalitions, in what Machado and Miguel (2011) describe as a programmatic dimension "connected to basic political values, which take into consideration the formation of alliances between parties [...] according to their categorization by ideology." They emphasize that a coalition is much more coherent the greater the affinity of the ideological of its members.

Therefore, it is reasonable to consider that ideological positions can be withdrawn from ideological positions and that the systematic verification of these positions is able to reveal ideological affinities. It is also reasonable to associate such affinities to a degree of coherence, or identity, within party coalitions. And why not say a degree of loyalty. In other words, party loyalty would not only be characterized by the act of desertion, but could be measured and monitored over time by what parliamentarians say in their pronouncements, with monitoring being either of the individual or a group of individuals.

The main problem resides in the difficulty of carrying out this systematic review, since it demands great data (texts in speeches) collection and analysis efforts. An analysis that is not restricted to content, for it seeks more than to simply extract meaning from the discursive text, aiming instead to understand it alongside the history of events. Based on the records of speeches made in the Plenary of the Chamber of Deputies, the manual execution of the verification process would be time-consuming and imprecise, in addition to not being able to contemplate all pronouncements.

Alternatively, the capture and management of the knowledge implied in the parliamentary pronouncement can be improved using computational resources and advanced linguistic processing techniques. Political scientists have used automatic content analysis on a diverse set of texts. This includes media data files (YOUNG; SOROKA, 2012), parliamentary speeches in legislatures around the world (QUINN *et al*., 2010), statements by the president, the legislator, and the party (GRIMMER, 2010), treated (SPIRLING, 2012), political science articles and other political texts.

In this sense, this study uses Natural Language Processing (NLP) techniques as a tool to treat and aid the analysis of parliamentary discourse, and as a useful resource in establishing connections between facts, quantitative data, and final results of the Legislative Process. Based on the preparation of data through NLP, this research innovates when proposing a textual similarity indicator, hereby called Ideological Identity Index, based on Chi-Square distribution (X^2), to measure the convergence degree of the pronouncement of deputies that belong to a certain parliamentary bloc or coalition. This measurement allows the investigation, to some extent, of the solidity of political alliances or, by extension, party loyalty.

The article is divided into seven sections in addition to this Introduction: the second

section discusses the concept of political-partisan ideology and the formation of coalitions; the third presents Natural Language Processing techniques; the fourth develops the mathematics to assess the similarity of discourse; the fifth proposed the Ideological Identity Index; the sixth discusses results and main findings; the seventh brings the final considerations.

2 Political-Partisan Ideology and Coalition

There is no intention here to discuss the concept of "ideology" in its theoretical-philosophical broadness, only to establish a pragmatic reference, a methodological resource with the purpose of identifying events based on parliamentary discourse, which can reveal a form of affinity, party loyalty and/or discourse coherence capable of characterizing a coalition.

The first theory about the process of coalition formation, developed by the political scientist William H. Riker, did not presume ideological proximity (RIKER, 1962). Maybe because the term ideology is understood by many meanings, sometimes divergent, this association was only proposed years later by Robert Axelrod (AXELROD, 1970).

Arguing that a single definition of ideology would be useless, Eagleton (1997) presented a list of the most common, from which the following stand out: set of action-oriented beliefs; body of ideas characteristic of a particular group or social class; ideas that help legitimize a dominant political power; identity thinking; the discourse and power set. According to Cancian (2007), the term "ideology" is used with the goal of "describing the set of ideas, values or beliefs that guide the perception and behavior of individuals about several social subjects or aspects," for example, "the opinions and preferences that individuals have regarding the political system in force, public order, government, laws, the social and economic conditions." For both authors, the sense of unity, a common guideline or identity, which is also related to the meaning of uniting, prevails in the concepts.

Machado and Miguel (2011), when proposing a typology for party coalitions, simplified the use of the concept of ideology by assuming that "stripped of its more complex (and more controversial) meanings" the word ideology "remits solely to the position in the left-right axis." Rodrigues (2009, p. 27), in his study on party caucuses, discusses the ideological logic of coalition under two angles: on one hand, it is believed that coalitions receive negative evaluation from public opinion because they unite ideologically and programmatically discrepant parties, in a scenario with high migration between parties; on the other, it is defended that most coalitions obey the logic of ideological affinity.

Reisman (2016) highlights that the formation of political coalitions is temporally located in the pre-electoral moment, when strategic alliances for majority and proportional elections are negotiated between political parties. When studying the electoral coalitions led by the PT in presidential campaigns, the author found that, even if the ideological formation criteria seem to be dominant, this supposed ideological convergence did not prevent the

movement of coalitions from left to center, showing that ideology submits to electoral logic. Such coalitions went through a drastic reduction in ideological weight and adopted generic and abstract electoral programs. In its pragmatic aspect, these coalitions abandoned their defense of controversial issues and approached the center of the left-right axis.

The phenomenon observed in the Reisman (2016) study, present in the overall political scenario, is probably the motivating factor of Constitutional Amendment No. 97, of October 4, 2017, which prohibited party coalitions in proportional representation elections (BRASIL, 2017), maybe to remove the electoral logic from the process of forming coalitions, stimulating ideological convergence in the conception of programs.

Therefore, simply put, this study proposed to use the concept of "ideology" with the complement "political party," representing the set of political guidelines with which a party is committed. Even if this scenario changes over time, it should reflect the party's commitment in each time frame. Based on this denomination, "partisan-political ideology," and under the perspective of the formation of party coalitions, one could expect that these associations take place, in principle, between parties that share the same political orientation, or at least most of them. Assuming this hypothesis as quite probable, we could suppose that two parties of a coalition would belong, or should belong, to the same partisan-political ideology.

Based on these aspects, two premises can be established to conduct this study:

1. in principle, a coalition should be based on the proposed concept of partisan-political ideology; therefore, a high degree of similarity is expected between the parties of the coalition.

2. an indicator for measuring ideological identity of a coalition must be able to capture the behavior of an entire group of parties at given periods.

Therefore, the main challenge consists of measuring the degree of similarity between the many parliamentary discourses of a coalition, which requires large-scale text content processing in the order of hundreds or thousands, a capacity only possible through computational resources.

The next two sections describe the technical-methodological tools used in this study to build an ideological identity indicator. The bag-of-words technique and the Chi-Square statistical method are presented with the following purposes: to prepare the text collections for computational treatment; to verify if specific two-word groups compose a language's own meaning; and establishing similarity metrics between text collections. The concept of *corpora* is used to refer to an ample collection of systematically organized digital texts on which a linguistic analysis is based, and that of *corpus* (singular of *corpora*), to a subset of texts extracted from the corpora (VYATKINA; BOULTON, 2017).

3 Natural Language Processing

Natural Language Processing is a sub-area of Artificial Intelligence and Linguistics that studies the problems of automatic comprehension of natural human languages. Grimmer and Stewart (2013), when studying political texts, recognize the benefits of NLP and how much automated processing reduces the costs and efforts of analyzing large text collections. They emphasize, however, that automated methods do not replace careful thinking and attentive reading, in addition to requiring extensive and specific problem validation. They argue that, for the automated text processing to become a standard tool for political scientists, researchers in the field must contribute to the creation of robust ways to validate methods.

Zhang *et al*. (2009) state that the ability to express the meaning of a work depends on the other words that follow it. When a word appears accompanied by a set of terms, the chances of this set having a relevant meaning are higher. This means that not only the word, but also contextual information is useful for information processing.

Silva and Souza (2014) emphasize that the text is not a simple group of random words, but that the order in which they are placed is what produces meaning. The study of the co-occurrence of words can indicate if they are directly related by compositionality or affinity, or, indirectly, by similarity. Therefore, the basis of empirical linguistics consists in finding the meaningful dependencies between terms, based on the frequency of observed co-occurrences. These adjacent terms are called n-grams, where *n* is the number of terms.

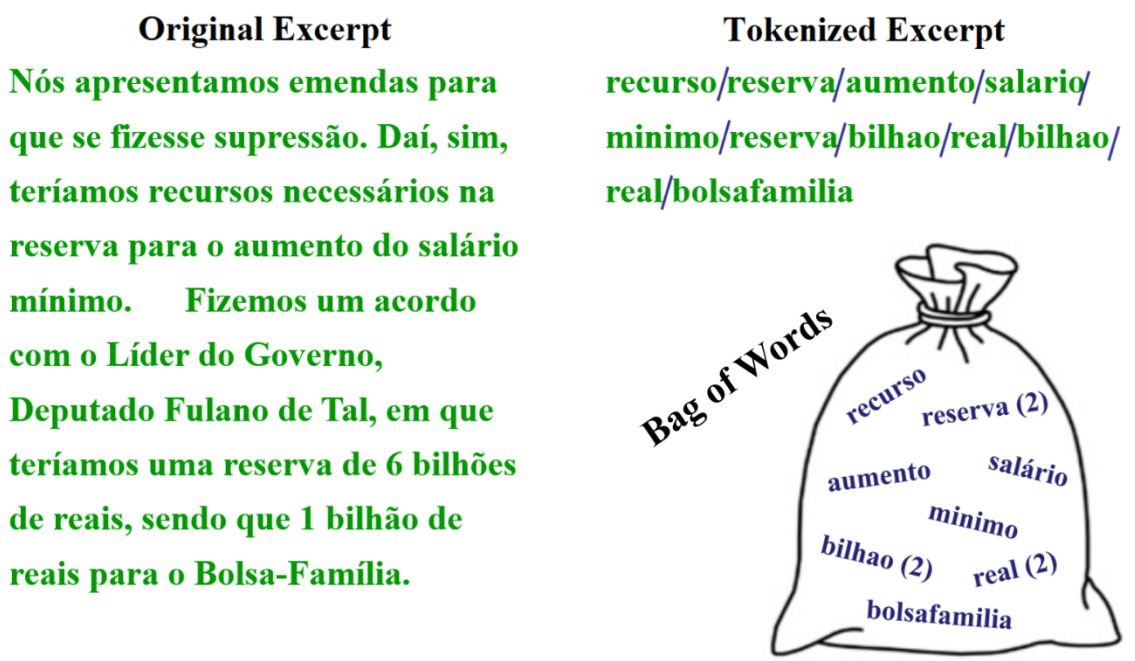
The n-gram language models were developed in the field of Statistics by the Russian mathematician Andrey Markov (1856-1922) to recognize statistical patterns of language use based on strings, known as Markov strings. Wang and Liu (2011) point out that many works predominantly focus on the identification and extraction of n-grams. This process is called tokenization, or segmentation of words, and consists in the task of reducing a text into minimal units called tokens, where each token can be a word, or set of words, a number, a punctuation mark, or another structure belonging to language (MANNING; SCHÜTZE, 1999).

The complexity of tokenization varies according to the complexity of the language itself, with the definition of the problem to be studied and with the model chosen to apply the NLP. Tokenization usually includes the following steps: converting the text characters to lowercase; removing the numbers, punctuation, plurals, diacritical signs, *stop words*[3] and blank spaces; and word segmentation. The final product of tokenization is a bag-of-words, with the respective frequency counts of terms. The bag-of-words approach, the order in which words are arranged in the texts is not considered, and the search for information only considers the estimated frequencies. Figure 1 illustrates an excerpt of the speech before and after the

[3] *Stop words* are words with little meaning in some NLP applications, such as information recovery and classification, which means that these words are not very discriminatory. Articles and pronouns are usually classified as *stop words*. The idea is to simply remove words that are very frequent in all documents of the text collection. However, the *stop word* list should be carefully chosen because it is determinant in reaching the goals.

tokenization process. The significant reduction of the text is mainly due to the set of *stop words* used.

Figure 1 – Tokenization applied to an excerpt of a parliamentary speech: bag-of-words



Original Excerpt (English version):

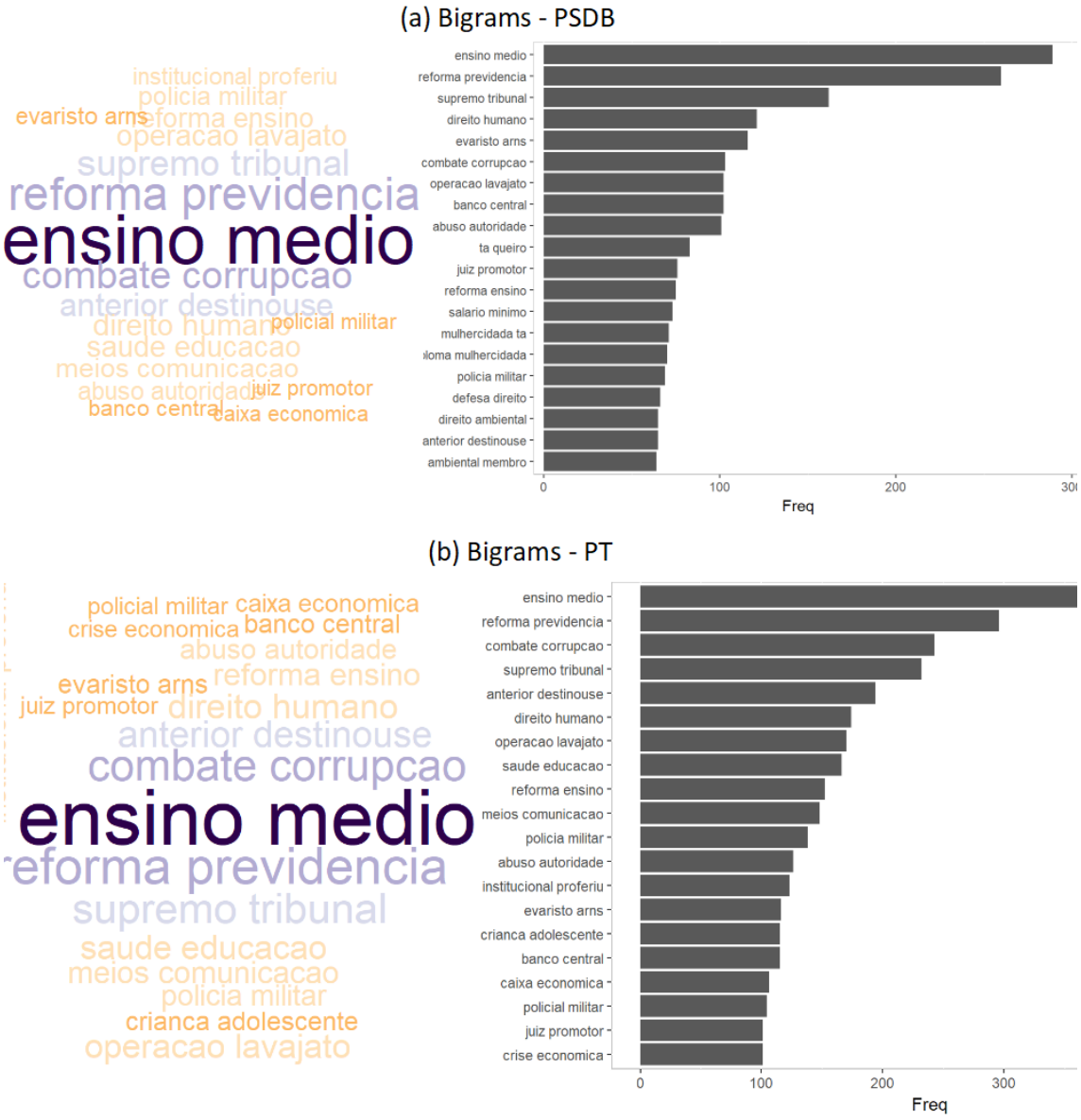
We present amendments to make suppression. Then, yes, we would have the necessary resources in reserve for the increase in the minimum wage. We made an agreement with the Government Leader, Deputy Fulano de Tal, in which we would have a reserve of 6 billion reais, with 1 billion reais for the Bolsa-Família.

Source: Own preparation.

The tokenization process can be adjusted to divide the text in bigrams, meaning two-word tokens that occur consecutively in the text. The bigrams are commonly used as the basis for statistical analysis of texts and help estimate the conditional probability of a word, given the preceding word (JURAFSKY; MARTIN, 2008). Bigrams are richer in meaning than the ascertainment of simple words and are more useful in extracting information from text.

Figure 2 illustrates the set of bigrams extracted from speeches given by deputies of the PSDB and PT parties, in 2016, presented in the form of word clouds and frequency charts. These resources constitute a heuristic analysis method that, by itself, does not allow the generalization of the answer to a research, but points out clues to what should or not receive more attention in a set of texts.

Figure 2 – Word cloud and absolute frequency bar chart: (A) bigrams of PSDB speeches in 2016; (b) bigrams of PT speeches in 2016.



Glossary:

abuso autoridade	abuse of authority
banco central	central bank
caixa economica	federal savings bank
combate corrupcao	fight corruption
crianca adolescente	child and teenager
crise economica	economic crisis
defesa direitos	defense of rights
direito ambiental	environmental rights
direito humano	human right
ensino medio	high school
evaristo arns	Evaristo Arns was a brazilian Franciscan friar
juiz promotor	judge prosecutor
meios comunicacao	media
opercao lavajato	Operation Car Wash was a criminal investigation by the Federal Police of Brazil
policia militar	military police
reforma ensino	education reform
reforma previdencia	pension reform
salario minimo	minimum wage
saude educacao	health and education
supremo tribunal	Supreme Court

Source: Own preparation.

When first looking at Figure 1, going against what common sense might expect about pronouncements from openly opposing parties, one has the impression that the speeches are similar and address the same themes, with slight variations. However, although important for the maturity of the investigative process, visual inspection is not enough to reach conclusions, with more robust approaches through statistical tests that will be detailed in the next section being necessary.

4 Similarity of the speech by comparing bigrams as *collocations*

The task of measuring how much a *corpus* of discourses resembles another through a computational method should preserve, as far as possible, the meaning of the texts. Although the bag-of-words technique does not have the same semantic evaluation function, since it disregards the order in which words are arranged in the text, the division into n-grams with n being greater than 1 ($n > 1$) results in sets of terms whose order corresponds to that of the original text. Therefore, these n-grams carry more semantic meaning than the isolated terms, making it reasonable to assume, for example, that the comparison between speeches based on bigrams is more effective than that based on isolated terms.

This assumption gains even more strength if we are able to identify the bigrams whose terms do not occur together by chance, but due to a dependence ratio that expresses a conventional way, in the language, of saying things, meaning it is inserted in a semantic context. Expressions of two or more words that correspond to a conventional way of saying things are called *collocation* (MANNING; SCHÜTZE 1999).

The statistical Chi-Square (X^2) statistical test can be used to assess whether two words constitute a *collocation*, verifying the dependence between them with the help of a contingency table (KILGARRIFF; ROSE, 1998). In essence, this test treats each word as a categorical variable and assumes the condition of independence of the events, meaning when the words are together by chance, and therefore do not constitute a *collocation*. By definition, two events are independent when the probability of both happening together is equal to the product of each event happening individually:

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

So, the expected value for the joint occurrence of two independent events is given by

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

where N is the total number of events.

Thus, knowing the expected value of independent events, the X^2 statistic is determined by comparing the observed frequencies with the expected frequencies. When there is no statistical difference between the observed frequencies and expected frequencies, it is concluded

that the events are independent. The hypothesis under the condition of independence of the bigram terms are stated by:

H_0: There is no association between the terms; the bigram is not a *collocation*.

H_1: There is an association between the terms; the bigram is a *collocation*.

In practice, the frequency with which the bigram occurs together, the frequencies with which they form bigrams with other words and the frequency of bigrams formed without these words are computed. Table 1 shows the distribution of the terms "high" and "school" (without diacritical signs, after tokenization) for the analysis of the bigram "high school", where w_1 and w_2 refer, respectively, to the first and second word of the bigram.

Table 1 – Contingency table: distribution of the terms "ensino" and "medio" for analysis of the bigram "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

Note: frequencies measured from PT speeches in 2016; "ensino medio" means "high school"; "oriente medio" means "Middle East"; "ensino superior" means "university education".

Source: Own elaboration.

Therefore, the X^2 statistic is measured in the following way from the contingency table.

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

where *i* represents the lines and *j* the columns of the table, O_{ij} is the value observed in the cell (*i*, *j*) and E_{ij} is the expected value. The values of E_{ij} are computed according to Equation 2, based on marginal probabilities, meaning, from the total of lines and columns converted into proportions. For example, for cell (1, 1), the expected value E_{11} is provided by the product between the probability of "high" being the first term of the bigram ($P(A)$), the probability of "school" being the second term of the bigram ($P(B)$) and the total of bigrams (N) that exist in the *corpus* under analysis:

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

Applying Equation 3 to all cells in Table 1, we reach $X^2 \approx 28.75$ and $p \approx 0.0000000821$. which means that H_0 should be rejected, meaning the words "high" and "school" do not randomly occur together. Therefore, they are dependent and constitute a *collocation*.

Once the *collocations* of two *corpus* are identified, the similarity between them can be measured, once again, through the X^2 statistic. Kilgarriff and Rose (1998) proposed this approach in a study where they compiled a table of *n* lines by 2 columns (with $n = 500$), in which each column corresponded to the total of most frequent words in common for both *corpus*. In this study, the Kilgarriff and Rose (1998) method was extended by bigrams validated as *collocations*. Table 2 contains a schematic representation of this conception, using the *corpus* of the speeches of the PT and PSDB parties in 2016.

Table 2 – – Most frequent collocations common to the *corpus* of discourses of PT and PSDB in 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	....	....	....

Note: "ensino medio" means "high school"; "reforma previdencia" means "pension reform"; "combate corrupcao" means "fight corruption"; "supremo tribunal" means "Supreme Court"; "direito humano" means "human right"; "opercao lavajato" was a criminal investigation by the Federal Police of Brazil; "saude educacao" means "health and education"; "reforma ensino" means "education reform"; "meios comunicacao" means "media"; "policia militar" means "military police".

Source: Own elaboration.

For each of the most frequent *n collocations* in Table 2, one must calculate the expected occurrences (expected value) in each *corpus*, aiming to verify if both can be considered random samples of a same population. Given the N_1 and N_2 sizes of *corpus* 1 and 2, the O_{i1} and O_{i2} frequencies of the *i*-thimal *collocation* in *corpus* 1 is given by the probability of a *collocation* happening in the total set (*corpus* 1 plus *corpus* 2) multiplied by the size of *corpus* 1 (N_1)

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

and, respectively in *corpus* 2, by

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

Knowing the observed and expected values, then X^2 can be determined by Equation 7,

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

where *c* represents the respective *corpus*. The null and alternative hypotheses can be formulated as follows:

> H_0: The observed frequencies are not different from those expected, meaning there is no difference between the frequencies of the groups.
>
> H_1: The observed frequencies are different from those expected, therefore, there is a difference between the frequencies of the groups.

Therefore, accepting H_0 means that there is no difference between the *corpus*.

Once the computational and statistical tools are defined, the next section applies the concepts explored in the construction of the proposed similarity indicator, based on the concept of partisan-political ideology.

5 Ideological Identity Index

From the proposal of Kilgarriff e Rose (1998) extended to bigrams, a first similarity test was performed for combinations of the political parties PT, PSDB, PMDB[4], PSOL, PCDOB and PTB, whose *corpus* of the year 2016 were compared two-by-two through the respective *collocations* (Table 3).

Table 3 – X^2 metric for corpus similarity: PT, PSDB, PMDB, PSOL, PCdoB and PTB political parties, year of 2016.

Party 1	**Party 2**	**Bigram (*collocation*)**		
		X^2	***p***	**H_0**
PT	PSDB	2,386.57	7,24e-302	rejects
PT	PMDB	1,836.77	1,67e-196	rejects
PT	PSOL	4,093.27	0.00	rejects
PT	PCdoB	4,217.40	0.00	rejects

[4] In the period corresponding to the structuring of the *corpora*, the party also went by the acronym PMDB, later changed to MDB.

PT	PTB	4,390.26	0.00	rejects
PSDB	PMDB	165.68	1.00	accepts
PSDB	PSOL	1,679.38	1,91e-221	rejects
PSDB	PCdoB	1,763.91	2,38e-238	rejects
PSDB	PTB	3,289.55	0.00	rejects
PMDB	PSOL	1,919.91	7,75e-277	rejects
PMDB	PCdoB	1,994.99	1,83e-292	rejects
PMDB	PTB	3,113.44	0.00	rejects
PSOL	PCdoB	16.83	1.00	accepts
PSOL	PTB	708.59	8,29e-32	rejects
PCdoB	PTB	648.03	8,91e-23	rejects

Source: Own preparation.

Table three shows the similarity of the *corpus* in the PSDB/PMDB and PSOL/PCdoB combinations. In a preliminary analysis, this seems reasonable in a year where President Dilma's *impeachment* took over the political scenario, and in which parties such as PSDB and PMDB came together for this cause, while others, such as PSOL and PCdoB took the opposing stance. It is also observed that PT's discourse is not like that of any other party, maybe reflecting a break in the governing base in face of the countless interests that surrounded the *impeachment* theme.

This first experiment showed to be coherent when comparing the *corpus*, revealing the ability to explain the events, which inspired the construction of an indicator meant to measure the degree of similarity of the discourse of a group of kin parties, called **Ideological Identity Index.** The idea consists of **determining the mean probability of the X^2 statistics found in the comparison of the party speeches that belong to a certain group, taken two-by-two**. Due to the assumed presumption of political-partisan ideology, it should be expected that this indicator has a value that is remarkably close to 1 (one) when measured in parties of the same coalition. By analogy, when applied to parties with opposite political orientations, it should yield values close to zero.

This study considered the speeches given in the short[5] and long address[6]. The choice of these phases in the ordinary plenary session is because they favor an ideological debate, less influenced by the pressure of the votes and heightened political moods.

Regarding this, Moreira (2016) identified that the speeches given in the small address did not concentrate on themes, "with indications that such strategy is especially guided by the

[5] The first part of the plenary's ordinary session lasts 60 minutes at most and is intended for communications by previously registered parliamentarians (CÂMARA DOS DEPUTADOS, 2020, p.48).

[6] Plenary session phase that follows that of the short address with a non-extendable duration of fifty minutes. Intended for the speech of registered parliamentarians where each speaker has up to twenty-five minutes to discourse, including any word that is granted for other reasons (CÂMARA DOS DEPUTADOS, 2020, p. 49-50).

number of speeches given and by the party's ideology." The author adds that "there are differences in thematic focus among the parliamentarians" and that "these differences are influenced by the percentage of votes received, the ideology of the party by which the parliamentary was elected."

This phenomenon may be due to the freedom that the parliamentarian has during the short address, where they can freely speak of any theme for up to 5 (five) minutes. Also according to Moreira (2016), "many deputies spend their entire term without speaking of votes and committees, but almost all speak at the short address," showing that in legislatures 51 to 54, 88% of the deputies made at least one speech during the small address, with an average 59 speeches per deputy. Likewise, the long address was included because it has two similarities with the short address: the deputy manifests the intention to speak by registering and there are no topic restrictions.

6 Results and discussion

Three distinct measurements of the Ideological Identity Index were taken, year-to-year, from 2001 to 2015. The first only considered the parties belonging to the governing coalition (or governing base), the second only considered the parties that do not belong to the governing coalition (hereby defined, in general, as opposition), the third estimated the similarity of the speech between parties that belong to the governing coalition and those that do not. The data on governing coalition compositions of each year were extracted from the Legislative Data Base of the Brazilian Analysis and Planning Center - Cebrap (CEBRAP, 2019). The full content of the speeches given in the Chamber's plenary of said period was downloaded from the Open Data Portal of the Chamber of Deputies (CÂMARA DOS DEPUTADOS, 2017). In total, more than 150 thousand speeches were read and processed, according to data and programming codes in Language R available in a public repository (CEFOR, 2018).

A similarity matrix like that of Table 4, which shows the values estimated for the year 2003, was generated for each year of the period. The parties written in capital letters correspond to the governing coalition.

Table 4 – Similarity matrix for speeches in 2003 $\mathbf{i_{coa}}$ is the ideological identity index for the governing coalition; $\mathbf{i_{opo}}$ is the ideological identity index for the opposition; $\mathbf{i_{dif}}$ is the ideological identity index when distinct groups are compared.

	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

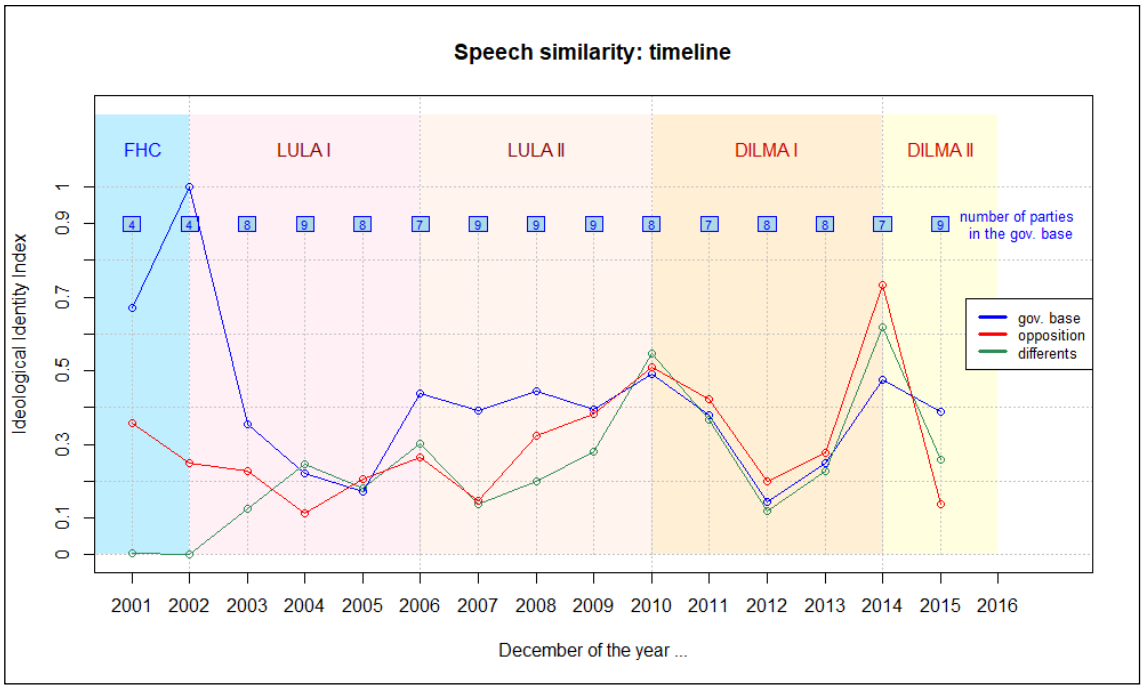
PPS	0.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.57	—	0.17
PR	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.43	—	0.17
prona	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.74	0.00	0.00	0.00	1.00	—	0.15	0.13
PSB	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	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
														Mean	**0.36**	**0.23**	**0.12**

Source: Own preparation.

Note: the parties written in capital letters correspond to the governing coalition, the crossover cells represent the probability of X^2 statistic resulting froom the comparison of the respective parties, two lines are highlighted as an example, one for the coalition (PCdpB) and another for the opposition (pmdb), the cells in blue draw attention to the terms used in the respective row to estimate the $\mathbf{i_{coa}}$; the cells in salmon highlight the terms used in the respective row to estimate the $\mathbf{i_{opo}}$; the cells in green highlight the terms used in the respective row to estimate the $\mathbf{i_{dif}}$.

Figure 3 shows the results found for the indicator in the period between 2001 and 2015, highlighting the respective governments at the time.

Figure 3 – Ideological Identity Index from 2001 to 2015.



Source: Own preparation.

It is possible to identify a coherent relationship between the results in Figure 3 and facts that that occurred during each phase of government, making it possible to assess the analytical potential of the proposed indicator. A first pattern observed is that of the rise in the governing base's indicator in electoral years (2002, 2006, 2010 and 2014), followed by a fall in the year

following the election. This behavior tends to confirm Reisman's (2016) argument that ideology submits to electoral logic. A similar movement is observed in the opposition's indicator, except for 2002, a year when Lula would be elected for his first term as President. On that occasion, when Lula started to grow in the opinion polls, a climate of insecurity spread in the political and economic fields, under the suspicion that the country could go bankrupt. Lula saw himself forced to sign a text that became known as Letter to Brazilians (SILVA, 2002), in which he stated he was willing to discuss a financial crisis response agenda with President Fernando Henrique Cardoso (FHC), which displeased left-wing sectors and politicians from the PT itself, the opposition party at the time, a fact that may be reflected in the indicator.

It is important to emphasize that in the experiment of Figure 3, the term opposition was used consciously in a broad sense, with any party that does not belong to the government coalition being classified as opposition, which is an accentuated simplification. Therefore, it is not a matter of an opposition coalition, but of political parties who are not in the government's supporting base, from which an ideological convergence should not be readily expected. To mention just one example, this criterion makes PPS appear in the same package as the PT in 2002. However, that was the year when the presidential candidate for the PPS, Ciro Gomes, was perhaps Lula's most emphatic critic. Anyway, this convergence appears in other electoral years.

Still in the FHC period, a strong convergence of the base is seen in 2001, and a unique situation of total convergence in 2002. It is worth noting that, at this time, the coalition was comprised of only 4 (four) parties, practically half of the amount that would come to comprise the bases of subsequent governments. It is natural to accept that reaching an identity is easier between 4 (four) parties than 8 (eight). When the differences (base and non-base) are compared, no identity traits are found.

In the Lula I period, the first two years show a drop in base and opposition, maybe due to the fact that a party that had always been opposition (PT) was now in charge and would encounter resistance in its own base (with eight parties in 2003 and nine parties in 2004) when supporting neoliberal policies, while, on the other hand, the party that led the country in the last 8 (eight) years (PSDB), now undertook a position of programmatic opposition, with a lesser degree of opposition, but with the possibility of victories in certain matters, as stated Bezerra (2012). This also explains the growth of the indicator for separate groups, since, according to the same author, parties such as PSDB and EM occasionally adopted the strategy of collaborating with the executive branch. In 2005, the base indicator registers the lowest value in the period, surrounded by accusations of corruption that triggered the governing coalition. In that same year, the opposition's indicator grew, as if resulting from the opposition's recovery of identity, influenced by the anti-corruption discourse. Finally, in 2006, the electoral logic prevails and both base and opposition sharpen the discourse in a growing movement of both indicators, with prevalence for the base indicator.

The Lula II period shows some stability in the base indicator, with a peak in 2010, which was an election year. Once the suspicions regarding the first Lula government had been overcome, the impacts of the Mensalão had been dampened, in face of a scenario of growth recovery and a stable economy, all associated with the President's political skill, it seems reasonable that there were no great variations in identity in the base's speech. After registering a minimum in 2007, the opposition indicator has a linear ascendancy until 2010. The indicator showing those that are different points to the similarity between speeches of the base and opposition in the election year, which is consistent with Lula's high popularity scenario (≈ 87%) and high government approval rate (≈ 80%). It would not be prudent for the opposition to base its speech on criticism towards Lula, which is why the PSDB's electoral campaign motto was "Brazil can do more," with the frequently repeated expressions "we are going to do more" and "we can do more and better" in their electoral programs.

The first year of the Dilma I period repeats the natural downward trend in ideological identity indexes. But the fall in the base indicator was not simply a natural trend. The international situation had become economically unfavorable, with low global growth, also affecting the Brazilian economy. In a scenario like this, the executive leader's political articulation abilities are a fundamental requirement for overcoming difficulties and approving the necessary measures. President Dilma proved to lack capacity, however, in the field of political skill, marked by her inability to dialog with political actors. For many times, she refused to talk to deputies and senators from the party itself, which led the government to lose votes and space to set agendas in the National Congress. In 2012, even without major social program launches or robust numbers in the economy, the government's popularity, and Dilma's assessment broke records, which probably reinforced the President's authoritarian profile. In that same year, the base indicator reached a level below that of 2005, the year of the Mensalão. The opposition indicator also dropped, a little less pronounced. As of 2013, the three indexes return to that of the election logic, in which the opposition indicator shows significant growth, reaching a maximum value in 2014, a reflection of an intense electoral dispute in which Dilma's reelection took place due to a difference of only 3.3% of valid votes.

In the only year measured for the Dilma II period, once again, the drop in indicators is confirmed. This phenomenon possibly results from natural adjustments motivated by new mandate settings at the municipal, state, and federal levels.

7 Final Considerations

The proposed Ideological Identity Index produced results consistent with facts that marked Brazilian political history in the period between 2001 and 2015, suggesting that the index has analytical usefulness in explaining these facts, which is a finding of this research.

In general, it can be said that the indicator constitutes an objective tool for discourse

analysis and that, under the conditions established in this study, it allows investigation of the behavior of political alliances, and may constitute an alternative to party desertion for measuring party loyalty. It can also be used to confront what the parliamentarian says and how they vote, meaning if the parliamentarian's vote is coherent with their speech.

In this sense, this work opens another path the NLP and quantitative methods to consolidate themselves as techniques for analyzing parliamentary discourse. In a field full of uncertainties, such as Political Science, new directions are always welcome to clarify issues.

References

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.