AUTOMATED CONTENT ANALYSIS METHODS APPLIED TO PARLIAMENTARY SPEECHES

Ricardo Modesto Vieira[1]

Abstract: Through substantially reducing time and cost of doing text-as-data research, automated content analysis methods have made feasible to research large collections, such as parliamentary speeches, quantitatively. This article identifies the conceptual framework of this new frontier of knowledge, describes a typical text-as-data project, presents the main methodologies and their applications to parliamentary speeches, shows the key challenges and best practices in this area of research, and finally suggests how to apply these methods to Chamber of Deputies' corpora. The intention is not to exhaust the range of methods, techniques and models; but to provide a guide for applying appropriately these methods to parliamentary speeches and other Chamber of Deputies' corpora.

Keywords: automated content analysis; text-as-data; analysis of parliamentary speeches.

INTRODUCTION

The "big data" revolution and artificial intelligence provided great opportunities for Chamber of Deputies to save labor, such as through automatic indexing; generate strategic information, such as the ideological positioning of the parties; and provide data for important discoveries, such as researchers Moreira (2016), Izumi (2017) and Schwartz (2018), through automated content analysis applied to parliamentary discourse[2]. This article aims to demonstrate how the speeches can be analyzed quantitatively with the aid of computational tools, to facilitate the organization and understanding of this vast collection, as well as generate indicators and strategic information for the Chamber of Deputies, their Executive Officers and their Leadership.

Representatives express their opinions, defend their positions and make propositions in words. According to the empirical approach adopted, what matters is the likelihood of 'common' and 'unusual' words and expressions appearing as linguistic events (MANNING; SCHUTZE, 1999, p. 7). Thus, through the frequency with which these words and expressions manifest themselves, it is possible to describe the use of language in the Chamber of Deputies. Of course, such a statistical approach will always be doomed to various errors and will never replace the

[1] Master of Political Science from San Diego State University, USAs. E-mail: ricardomodestovieira@gmail.com

[2] This is not about the French lines of Discourse Analysis - which is referenced by Michel Pêcheux and Dominique Maingueneau – or Content Analysis - which is referenced by Laurence Bardin, but the US line of Natural Language Processing, whose references are George Kingsley Zipf and Alan Turing. Thus, the empiricist view of these two authors differed from Chomsky's rationalist linguistics. In this article, the focus is not on the meaning of discourse or its content, but on the use of automated quantitative techniques applied to a specific content: parliamentary speeches and debates. The term "speeches and debates" has been officially defined by the Chamber of Deputies as appropriate to refer to both the speaker's speech alone in the podium – as well as the Small and Great Expedient – as well as the discussion of matters on the Order of the Day.

human component. However, with the advancement of computational tools, automated content analysis methods have become useful tools for exploring a vast collection such as parliamentary speeches.

In this sense, ten different methods will be analyzed – four classification methods, including two topic models, three scheduling methods, and three text similarity methods – and their applications for Chamber of Deputies. All of these methods use bag-of-words as an assumption: what matters is how often certain words are found in the corpus or corpora. As there is no global method for automated content analysis (GRIMMER; STEWART, 2013), the choice was made for its use in the Political Science literature and for its opportunity for application to the Chamber of Deputies in view of the variability of possible purposes.

To understand these ten different techniques and their possible applications, it is first necessary to explain the fundamentals of natural language processing as well as the data preprocessing techniques that enable automated content analysis. After explaining the preprocessing as well as the description of the techniques and their applications, the most common challenges related to these automated analysis methods applied to parliamentary speech will be analyzed, as well as the best practices recommended in the literature to address these challenges. Finally, some possible applications of these methods to the Chamber of Deputies will be suggested.

BACKGROUND AND DATA PREPROCESSING

In an empirical approach, the human brain starts from associations, pattern recognition, and generalizations to learn the detailed structure of natural language. There is thus no innate faculty of language as a genetic inheritance. There is no difference between language proficiency — knowledge of language structure — and one's linguistic performance in the world. Since language is inseparable from its social context, what matters are the common patterns that occur in the use of language (MANNING; SCHUTZE, 1999, p. 5-6).

Through a statistical approach, it is possible to automatically learn these lexical preferences and the words and expressions that tend to group together and form their own lexical field. Because the meaning of words / expressions is related to the context in which they are used, this knowledge of lexical preferences can be explored to understand deeper semantic relationships (MANNING; SCHUTZE, 1999, p. 18-19).

Following this statistical approach, Zipf (1949) found that there is a relationship between the frequency of a word and its rank in a list of words in the same corpus[3]. This constant is useful as

[3] A set of texts is called a corpus. Several of these text collections are called corpora (MANNING; SCHUTZE, 1999, p. 6).

a rough description of the frequency distribution of human language words: there are some very common words, an intermediate number of medium frequency words, and many low frequency words[4]. Zipf also obtained empirical evidence for the tendency of words belonging to the same content to group together (MANNING; SCHUTZE, 1999, p. 24-25).

Thus, to reduce the complexity and size of vocabulary, as well as focus on what is usual and meaningful in the text, only medium frequency words are analyzed. Therefore it is necessary to remove the words of unnecessary content and infrequent — hose that appear in 99% or less than 1% of documents (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018, citando HOPKINS; KING, 2010; QUINN et al., 2010). The most common words usually do not generate meaningful content and correspond to conjunctions, prepositions, articles, pronouns and copulas.

When the order in which words occur in the text is not considered, the bag-of-words principle is assumed. (BLEI, 2012; GRIMMER; STEWART, 2013). The bag-of-words principle presents each document as a single vector with the length equal to the number of unique words in the text (ROBERTS et al., 2015; IZUMI; MOREIRA, 2018). To restore some importance to word order, one can use bigrams, trigrams (GRIMMER; STEWART, 2013) or *collocations*[5].

It is also important to emphasize the need to remove what in the literature is called *stopwords* (GRIMMER; STEWART, 2013; ROBERTS et al., 2015; CASAS; WILKERSON, 2017). Each context generates a different list of stopwords, that is, words, phrases, and expressions that are used very commonly in that particular context but do not generate meaningful content. Lauderdale and Herzog (2016), for example, withdrew all procedural speeches from the legislative process, such as the President's speeches, the reading of the minutes and agenda, Bureau elections, prayers and honors.

After the removal of stopwords, it is necessary to reduce word variability through *stemming* or *lemmatization*. Stemming is the reduction of the word to its radical by removing its end, as in plurals or verb conjugations. In fact, stemming is an approximation of a linguistic concept called lemmatization, which seeks to reduce words to their basic form and group them, using a more complex algorithm that identifies the origin of the word and returns only its lemma or root. (GRIMMER; STEWART, 2013; ROBERTS et al., 2015; CASAS; WILKERSON, 2017). In Portuguese, an adaptation of Porter's algorithm (1980) is used for stemming (IZUMI; MOREIRA, 2018).

After stemming, the individual occurrences of each word are called tokens (MANNING; SCHUTZE, 1999, p. 22) and the document content is finally ready to be converted to quantitative

[4] Because the line is too low for most low ratings and too high for ratings over 10,000, Mandelbrot obtained a more accurate relationship between rating and frequency using three other text parameters as variables (MANNING; SCHUTZE, 1999, p. 25).

[5] Collocations is a commonly used expression where the whole is greater than the sum of the parts. Any expression that people repeat is a candidate for a collocation. The two most common word collocation' patterns are "adjective noun" and "noun noun" (MANNING; SCHUTZE, 1999, p. 29-31).

data. Remember that before preprocessing it is necessary to obtain[6], encode[7] and process[8] the documents. The bag-of-words is the simplest and most commonly used method for turning each document into a single vector whose value is determined by the absence or presence of a token in the document, the frequency of these tokens, or the frequency normalized by the document size (DIERMEIER et al., 2012). The goal is to create a Document Term Matrix (DTM) in which each row represents a document and each column represents a unique token. Since each matrix cell denotes the number of times the token indicated in the column appears in the document indicated in the row, each document is represented by a unique vector (ROBERTS et al., 2015; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

METHODS AND APPLICATIONS

After obtaining, encoding, treatment and pre-processing parliamentary speeches, it is possible to use various methods of automated content analysis for different purposes — such as saving labor, generating strategic information and making scientific discoveries. These methods can be divided into three sets, according to the objectives and tasks they intend to accomplish: (1) classification; (2) staggering and (3) text similarity. The first proposes classifies texts or documents into known or unknown categories. The second proposes to estimate the location of actors using a scale. And the third proposes to measure the similarity and / or homogeneity of texts or parts of these texts.

There are also two different approaches: supervised and unsupervised methods. The main difference between these approaches is that, in supervised, it is necessary to specify the conceptual structure of the texts beforehand, while in the unsupervised one uses a model to find a low-dimensional abstract that best explains the observed documents, and only the number of categories need to be informed in advance (ROBERTS et al., 2015). The significance of this dimension that the method recovers in the case of parliamentary speeches depends on the political context, as preferences and motivations for speeches vary according to this context.

However, the results of the low dimensionality of indicators elaborated through automated content analysis have been recognized as an important feature of congressional decision-making,

[6] Documents can be obtained in several ways: (1) open data in JavaScript Object Notation (JSON), Extensible Markup Language (XML), Comma-Separated Values (CVS), etc.; (2) documents of various kinds converted to editable data by optical character recognition (OCR); (3) web scraping data methods in which the computer accesses web pages by copying and organizing their content; (4) online platforms such as Amazon's Mechanical Turk, and applications (APIs) that allow the request of selected content from a database (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

[7] Encoding is the way in which the computer translates individual and unique characters into bytes so that the text can be read by the machine (ROBERTS et al., 2015; IZUMI; MOREIRA, 2018). For Latin characters, UTF-8 (GRIMMER; STEWART, 2013) is used.

[8] Similar formatted documents are required. Identical formatting makes it possible to write a single script to extract more specific content from multiple documents at once (CASAS; WILKERSON, 2017).

especially when one considers that the use of words in political debate also varies according to topic debated[9] and the type of political or ideological scale estimated (LAUDERDALE; HERZOG, 2016). Thus, the low dimensionality in Congress implies, for example, that the registration of a representative's vote on one issue will be a good predictor of his choice of vote on another unrelated issue (DIERMEIER et al., 2012).

1.1 Classification Methods

Classification methods may be supervised and unsupervised. Supervised learning methods use the frequency with which words appear in text to classify documents into predetermined categories or to measure the extent to which documents belong to specific categories. The algorithm then "learns" how to classify documents into these categories using a training set. That is, the algorithm uses document characteristics to classify them into categories. Because there are clear statistics that summarize model performance, supervised methods are easier to validate (GRIMMER; STEWART, 2013).

Within supervised methods, those that previously require the identification of the words that separate the classes are called dictionary methods. Dictionary methods use the relative frequency of keywords to measure the presence of each category in text. Using a list of words associated with speech tone (annotated dictionary), as well as the relative frequency at which these keywords occur, it is possible to measure the tone of a document (GRIMMER; STEWART, 2013). However, sentiment analysis — another important area of classification research, where the goal is to sort text, ordinally (from negative to positive, for example) rather than categorically — can be employed using supervised and unsupervised methods (CASAS; WILKERSON, 2017).

Unsupervised mixed association models or topic models are a set of Bayesian generative models that encode the specific structure of the problem into a category estimate (GRIMMER; STEWART, 2013). In other words, they are a set of algorithms that aim to discover and identify documents with thematic information (BLEI, 2012). Algorithms analyze text words to find topics, how these topics are connected to each other, and how they change over time. Algorithms do not have information about these topics, and documents are not labeled with topics or keywords (BLEI, 2012). Since statistically a topic is a probabilistic function on words, to estimate a topic, models use word co-occurrence between documents (GRIMMER; STEWART, 2013).

Thus, these unsupervised methods use text characteristics without imposing predetermined categories, using only modeling assumptions and text properties to estimate a set of categories and simultaneously assign documents (or parts of documents) to these categories (GRIMMER;

[9] The political association for certain words depends on the debate in which those words were used – a word that implies a leftist position in one debate may imply a rightist position in another debate – but also some words are similarly used to denote the position in many debates.

STEWART, 2013). Interpretable topic distributions arise by computing the hidden structure that likely generated the observed document collection. Thus, the process that generates the topics defines a distribution probability in relation to observable and unobservable random variables (BLEI, 2012).

Classification methods and topic models can be used for a variety of purposes. These methods can be used to study legislative issues, topics and agendas, such as understanding how political agendas change over time (DIERMEIER et al., 2012). Supervised methods can serve for automatic categorization and indexing of documents, saving labor, as well as sentiment analysis and positioning of political actors. Unsupervised methods of classification can answer questions about the internal functioning of government, the influence of different political groups and actors on public policy formulation, political positioning of actors, and so on. (ROBERTS et al., 2015).

Recent practical examples include Moreira (2016), who researched the government-opposition relationship based on parliamentary speeches through the Expressed Agenda Model topic model and noted that the thematic emphasis of the representatives does not follow the government-opposition relationship found in roll-call votes. Izumi (2017), who estimated political positions in the Senate through the Naive Bayes classifier (sentiment analysis) applied to Senators' speeches; and Diermeier et al. (2012), which used Support Vector Machines (SVM) in conjunction with the roll-call votes' database to predict the ideological position of representatives as well as to measure the degree of intra-party cohesion in the Senate. Compared to Naive Bayes, Diermeier's results showed that SVM is superior for classifying ideological positioning.

1.1.1 Naive Bayes

Naive Bayes is one of the most widely used supervised classification methods. Although it starts from a naive assumption — the model assumes that words are independently generated for a given category (the naive assumption), when in fact word usage is highly correlated in any data set —, the model provides a useful alternative method for assigning documents to predetermined categories (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). For large collections, the Decision Tree classifier can be used in conjunction with Naive Bayes to increase accuracy (KOHAVI, 2011).

For the supervised method to work, one must first (1) construct a training set — in which there is (1.1) the creation of a manually coded scheme and (1.2) the random choice of sampling documents (as a general rule between 100 and 500 documents); then (2) apply the supervised learning method so that the algorithm "learns" — learning the relationship between characteristics and categories in the training set and then using it to infer labels in the test set — how to classify the documents in the categories using the training set; and (3) validate model output by comparing automated coding output to manual coding output. Supervised learning methods are much easier to validate, with clear statistics that summarize model performance. After these three procedures

it is possible to successfully classify the other documents (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018).

1.1.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised method that uses a text classification algorithm to extract the most indicative terms from conservative and liberal positions in legislative speeches and to predict the ideological positions of representatives. SVM is based on the principle of structural risk minimization of statistical learning theory. First, speeches by the 25 most liberal and conservative US senators were analyzed — using the DW-Nominate (voteview.com) scores, which measure ideological positioning based on roll-call votes — o train the classification algorithm (DIERMEIER et al., 2012).

In the training phase, one category was arbitrarily labeled as "negative" and the other as "positive". Because the data points in each category are separable by a hyperplane, there are two parallel hyperplanes where the distance between the points in each category is as large as possible. These data points are called support vectors, and the distance between the two parallel hyperplanes is called the margin. The task of SVM in the training phase is to find the two separation hyperplanes so that the margin is maximum. Then, the same methodology was used to investigate moderate senators, training the classifier in moderate senators (DIERMEIER et al., 2012).

1.1.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is the simplest model of topic or mixed-association grouped data. The idea behind LDA is that each document is a mix of various topics in different proportions. In addition to the bag-of-words assumption, there is the assumption that the number of topics is known and fixed. Technically, the model assumes that topics are generated first, before documents (BLEI, 2012; GRIMMER; STEWART, 2013).

For each corpus document, words are generated in two steps: (1) randomly choose a distribution on the topics; (2) for each word in the document (2.1) randomly choose a topic from the first step distribution and (2.2) randomly choose a word from the corresponding distribution over the vocabulary. The distribution that is used to draw the pre-document topic distribution is called the Dirichlet distribution, and the Dirichlet result is used to allocate document words for different topics. All documents share the same set of topics, but each document displays these topics in a different proportion (BLEI, 2012).

1.1.4 Structural Topic Model (STM)

STM is also a mixed-association topic model, but it provides a flexible way to incorporate metadata associated with text — such as when the text was written, where it was written, who

wrote it, author characteristics, and so on — as covariates in document analysis (ROBERTS et al., 2015). The inclusion of document metadata follows and extends the Dynamic Topic Model — model in which the likelihood of observing a topic changes over time — and the Expressed Agenda Model — model that includes information about document authors, assuming each author splits his / her attention to a set of topics. Covariates also allow metadata sharing depending on the frequency of topics, such as the likelihood of women speaking a specific topic or using certain words in relation to men. STM also distinguishes itself from LDA by replacing the Dirichlet distribution with a normal logistic distribution, as in the Correlated Topic Model, to estimate the correlation between topics. Thus, it is possible to draw a network of correlated topics for a structured topic model, using a theme as a topic predictor in a given corpus (GRIMMER; STEWART, 2013; ROBERTS et al., 2015).

1.2 Scheduling Methods

Scheduling Methods are used to locate politicians and parties in continuous ideological spaces (KLUVER, 2009; DIERMEIER et al., 2012; GRIMMER; STEWART, 2013; LAUDERDALE; HERZOG, 2016; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018). These methods are based on the assumption that the ideological trends of political actors determine what is discussed in the texts — the assumption of ideological dominance in speech (GRIMMER; STEWART, 2013). But this may not be true because, as Mayhew (1974) has shown, politicians regularly engage in non-ideological credit claims. Thus, scheduling methods will perform better if they are accompanied by methods that separate ideological and non-ideological statements (CASAS; WILKERSON, 2017).

The best known use of these methods is the Vote View (voteview.com) project. Because ideology is not directly observable, Poole and Rosenthal (1991) developed a two-dimensional spatial model (*D-Nominate scores*) to rank representatives by their roll-call votes. The first dimension represents the traditional left-right view associated with the role of government in the economy and income redistribution. The second dimension represents issues of state interference in private life, slavery, and subsequently racial and civil rights issues. The model can correctly classify 85% of individual voting decisions of each member of Congress (DIERMEIER et al., 2012).

Another popular application is the Manifesto Project, which uses manual decoding to classify more than 1,000 parties in 50 countries according to their political manifestos (manifesto-project.wzb.eu). In this case, the ideal point estimate is used to classify the parts as reference points on a left-right ideological scale (DIERMEIER et al., 2012).

In addition, Lauderdale and Herzog (2016) estimated the individual political positions of each US representative using the Wordshoal scheduling method, which combines Wordfish and Bayesian factor analysis. The results of this research also suggest that individual political

positions of representatives are accurate predictors of intra-party cohesion and dissident behavior in countries where electoral systems provide strong incentives for personal votes (LAUDERDALE; HERZOG, 2016). Finally, Pritoni (2014) analyzed the difficulties in measuring the influence of interest groups by extracting their political positioning and comparing it with legislative products through scheduling methods.

1.2.1 Wordscore

Wordscore is a supervised scaling algorithm and a special case of dictionary method. The first step is the selection of reference texts that define political positions in space as liberal and conservative. Reference texts (training) are used to generate a punctuation for each word. Punctuation measures the relative rate at which each word is used in reference texts. This creates a measure of how well the word separates liberal and conservative representatives. Then, word punctuation is used to size the remaining texts (GRIMMER; STEWART, 2013). That is, the second step is to generate scores for the words of reference texts based on the political position assigned *a priori* and weighted by the probability of observing it in a document. The same procedure can be used for the left-right ideological scale (IZUMI; MOREIRA, 2018).

Wordscore is based on a number of assumptions: (1) political positions are reflected in the relative frequency of words used within and between texts; (2) the meaning of words remains stable over time; (3) all words have the same weight in the estimation process; and (4) all words of interest are contained in the reference texts (KLUVER, 2013). It is still necessary to deal with the definition of the political dimension to be investigated and choose a set of reference texts with known political position estimates, preferably from an independent source (KLUVER, 2013). It is also important that reference texts use the same lexicon as the texts to be tested, covering the entire ideological spectrum and having a diverse set of words (IZUMI; MOREIRA, 2018).

1.2.2 Wordfish

Wordfish is an unsupervised algorithm that estimates the importance of words to discriminate political positions based on item response theory (IRT). The model assumes a Poisson distribution for word count and assumes that the probability of looking at one word in a document is independent of the position of the other words in the same document. Thus, it can be used to discover words that distinguish positions in a political spectrum (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). Wordfish does not require reference text in contrast to Wordscore. In this case, the researcher must define the political dimension to be analyzed, select the documents that deal with this political dimension and remove all text passages that do not refer to the investigated dimension (PRITONI, 2014). Therefore, careful validation is required to confirm that the intended ideological space has been identified (GRIMMER; STEWART, 2013).

1.2.3 Wordshoal

Wordshoal is a hierarchical factor model for speaking in legislative debates that combines two approaches into a single estimation strategy. The first is to limit the analysis to speeches on a single legislative topic, keeping the topical variation constant. The second approach is to combine many speeches on many legislative topics into one document for each legislator or party. Ou seja, no primeiro estágio o modelo usa a escala de texto existente Wordfish para mensurar a variação do uso de palavras em cada tópico separadamente. That is, in the first stage, the model uses the existing Wordfish text scale to measure the variation of word usage in each topic separately. In the second stage, Bayesian factor analysis is used to construct a common scale from the specific debate positions estimated in the first stage. Thus, the model presents the results of Wordfish-based estimates for each debate, and then uses these estimates as data for the second stage aggregation model to assess whether a representatives is usually on the right or left of another representatives in a set of debates on heterogeneous themes (LAUDERDALE; HERZOG, 2016).

More specifically, the model uses a one-dimensional Wordfish scale applied to a set of texts within a single political debate, thus maintaining a constant variation of topic-oriented word usage. This Poisson scale model applied to each debate results in a specific estimate of the debate on the relative position of each representative. Having estimated the position expressed for all representatives on a given topic, the model aggregates specific dimensions of debate that involve varying subsets of legislators into a smaller number of dimensions that include all legislators. Since this approach does not depend on varying the use of speech in any debate to estimate positions in a latent dimension of disagreement, it is possible to generate one (or more) latent general position for each legislator with great predictive power (LAUDERDALE; HERZOG, 2016).

The empirical assumption of Wordshoal is that political disagreement is more clear and consistently reflected in variation within the debate on word use than in variation in word use in various debates. One of its main innovations is that Wordshoal allows the meaning and discriminatory power of a given word to vary from debate to debate. Variation in the use of words between speeches is as much a function of the topic of debate as a function of the position a legislator takes. In addition, the method provides significant uncertainty estimates of the aggregate positions of legislators, considering the frequency with which representatives spoke and the consistency with which they expressed their positions in debates (LAUDERDALE; HERZOG, 2016).

1.3 Text Similarity Methods

Text reuse consists of discovering instances of similarity in language usage. The distinctive feature of text reuse algorithms is that they explicitly value word sequencing when judging document similarity (CASAS; WILKERSON, 2017). However, since corpus resemblance is inherently multidimensional (they will be similar in some way and different in others), a measure of similarity only makes sense when comparing two homogeneous corpus. Thus, similarity can only be interpreted in light of the homogeneity of the corpus. In this sense, the same measure can be used for similarity and homogeneity by comparing the distance between two corpus (within-corpus distance) (KILGARRIFF; ROSE, 1998).

Similarity methods can be used to compare documents as a whole or find small pieces of text between two documents, such as those found in legislation drawn from multiple sources. To study whether congressional bills reuse texts from other congressional bills, Wilkerson et al. (2015) used the Smith-Waterman algorithm to compare text strings of US Congress bills introduced since 1990. In addition to revealing standards about representatives who introduce similar projects in a legislature or between legislatures, the algorithm can be used to determine to what extent the language a legislator introduces matches that of other legislators (BURGESS et al., 2016).

Already Hertel-Fernandez and Kashin (2015) used similarity methods to trace the origins of propositions and unravel the influence of interest groups in the legislative process (CASAS; WILKERSON, 2017). In the same sense, Burgess et al. (2016) unraveled the reuse of texts in propositions, using ElasticSearch to limit the number of comparisons to detect the influence of interest groups on the proliferation of state legislations in the United States (Legislative Influence Detector Project: dssg.uchicago.edu/lid/).

Finally, similarity methods can also be used to position representatives and parties on an ideological spectrum. In this sense, Schwartz (2018) used the X^2 technique to test the similarity of Great and Small Expedient speeches — through the frequency of words and bigrams (collocations) in political party discourse sets (PT, PSDB, PMDB, PSOL, PCdoB and PTB) compared two by two.

1.3.1 Smith-Waterman

The Smith-Waterman algorithm is designed to find similar subsequences within long chains of DNA, but it can also be used to find excerpts of a document that are similar to excerpts of other documents by an alignment score based on three parameters: match of words (matching), mismatch of words (mismatching) and gaps. This algorithm is a good choice for comparing a relatively small number of documents, but it may take a long time to execute in a large corpus (BURGESS et al, 2016).

To solve this problem, Burgess et al (2016) used the ElasticSearch search engine, configured with the Lucene default scoring function, to classify documents for a given query and thus filter

the set of documents executed by identifying a subset of corpus documents most likely to contain text similar to that of the query document. Filtering has been shown to increase efficiency because the local alignment algorithm compares only the documents returned by the search module without sacrificing accuracy in document similarity tasks (BURGESS et al, 2016).

1.3.2 Cosine

As already shown, each document can be represented by a vector, the length of which is equal to the number of unique words in the text. Thus, it is assumed that the greater the similarity in the relative frequency of the words used, the greater the similarity of content between texts. Being two vectors "u" and "v", it is possible to calculate the similarity through the internal product between them, because the larger the internal product between them, the higher the frequency for the same words. As this measure is still problematic, the solution is to divide the internal product by the product of the vector lengths, which is mathematically represented by the cosine of the angle formed between the vectors "u" and "v (IZUMI; MOREIRA, 2018).

1.3.3 X^2

Kilgarriff and Rose (1998) presented a method for evaluating corpus similarity – Known-Similarity Corpora – and tested approaches commonly discussed in the literature: measures of cross entropy[10], Spearmen and X^2. For the size of the corpus used – a subset of the British National Corpus containing 300,000 newspaper and periodical words divided into 10,000 word pairs – the X^2 and Spearmen approaches performed better[11] than any of the cross entropy measures; between the two, X^2 surpassed Spearmen. For each of the most common words, the authors calculated the number of expected occurrences in each corpus if both corpus were random samples from the same corpora. Since a corpus is never a random sample of words, the difference in the frequency of each word between two corpora tends to increase, but does not increase in the order of magnitude, as with gross frequencies (KILGARRIFF; ROSE, 1998).

[10] Perplexity is, roughly speaking, a measure of the size of the set of words from which the next word is chosen, given the history of the words. Perplexity is used to evaluate how good a language modeling strategy is, considering the same corpus. Thus perplexity can be used to measure a property similar to homogeneity if the language modeling strategy is kept constant and the corpora are varied. With the language modeling strategy kept constant, cross entropy becomes a measure of similarity (KILGARRIFF; ROSE, 1998).

[11] The reliability test (called gold standard) of the methods was performed having as parameter the comparison of two sets of pairs by the coders (KILGARRIFF; ROSE, 1998).

Table 1- Summary of Automated Content Analysis Methods

Method	Technique	Description and Application
Classification	Naive Bayes	Supervised method that classifies documents into known categories from a training set. The Decision Tree classifier is used to increase accuracy in case of large collections. It can be used for automatic indexing, classification of most debated topics, sentiment analysis and ideological positioning of representatives.
	SVM	A supervised method, Support Vector Machines (SVM) is used in conjunction with the roll-call votes database to predict the ideological position of representatives, to correlate what they said and how they voted, as well as to measure the ideological consistency of parties and degree of intra-party cohesion in Congress.
	LDA	Unsupervised method, Lattent Dirichlet Allocation (LDA) classifies documents without having to specify categories in advance, but the number of categories must be entered.
	STM	Unsupervised method, the Structural Topic Model (STM) allows to enter metadata as covariates by combining Expressed Agenda Model and Dynamic Topic Model. It can be used for classifying most debated topics as well as understanding political positions, leadership patterns and influences in the legislative process.
Scheduling	Wordscore	Supervised method that politically positions documents in known dimensions from a training set. It can be used to position political and parliamentary parties on an ideological scale – as left-right or liberal-conservative – through their speeches.
	Wordfish	Unsupervised method that politically positions documents without having to specify dimensions in advance, but the number of dimensions must be entered. It can be used to predict political positions and position political parties and representatives on an ideological scale.
	Wordshoal	Wordshoal combines Wordfish and Bayesian factor analysis to estimate each representative's position relative to each other within the party and between parties based on his speeches. It can also be used to estimate the coherence of parliamentary positioning over time, as well as intra-party cohesion and dissident behavior.
Text Similarity	Smith-Waterman	Smith-Waterman is an algorithm that measures the reuse of parts of a text and is used to measure the similarity of excerpts between two texts. It can be used to identify the source of parts of propositions, reports, amendments and approved legislation. It can also be used to measure the influence of interest groups on legislative product.
	Cosine	The cosine technique assumes that the greater the similarity in the relative frequency of the words used, the greater the similarity between two texts. It can be used to identify how similar two propositions or amendments are.
	X^2	The X^2 statistic is used to measure similarity and / or homogeneity between corpus. The X^2 calculates the number of occurrences of the most common words in each corpus and measures the difference between the frequency of these words in the two corpus. Can be used to measure ideological identity between parties and party blocs based on representatives' speeches.

Source: The author (2018).

METHODOLOGICAL CHALLENGES AND GOOD PRACTICES

Since automated content analysis methods are not a substitute for human reading, careful validation of results is required (GRIMMER; STEWART, 2013), based on the replicability of results and rigorous manual coding, preferably with more than one coder. Thus, validation is a critical component of every text-as-given project (CASAS; WILKERSON, 2017). For supervised classification and scheduling methods, it is important to demonstrate that computerized classification replicates manual coding. A classifier can be refined through a codebook and manual coding iterations (GRIMMER et al, 2018). In unsupervised methods, however, there is no such gold standard, validation occurs as the parameters are adjusted to examine new results (GRIMMER et al, 2018). One way to do this is to examine the degree of cohesion and word distinction of each topic (ROBERTS et al, 2014; CASAS; WILKERSON, 2017). Unsupervised methods also require validation that the measurements produced correspond to the claimed concepts (GRIMMER; STEWART, 2013).

It is worth remembering that, besides being multidimensional, texts are more flexible than other types of variables, creating a wider range of potential text properties to be analyzed and validated (GRIMMER et al, 2018). Lauderdale and Herzog (2016), Roberts et al (2015), Diermeier et al (2012) and Kilgarriff and Rose (1998) also mention the issue of text multidimensionality. Since all text is multidimensional, it is necessary to choose one or a few dimensions (low-dimensional representation) to understand the corpus and make inferences. This limitation is inherent in all three sets of methods[12]. Of course, the problem of multidimensionality in the case of ideological position also runs into the difficulty of operationalizing complex concepts such as ideology, left-right, conservative-liberal, etc.

Therefore, it is important to validate results with real-world events and expected events (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017) by linking them, for example, with facts, quantitative data, and outcomes of legislative activity (SCHWARTZ, 2018).). Another good validation practice is to use different algorithms for the same purpose and compare whether the results are similar (CASAS; WILKERSON, 2017 citing QUINN et al, 2010, GRIMMER; KING, 2011, ROBERTS et al, 2014). In this sense, supervised learning methods can be used to validate or generalize the findings provided by unsupervised methods (GRIMMER; STEWART, 2013). There is also a trade-off between the degree of generalization of the concept and the validity of empirical indicators (PRITONI, 2014). It is tempting to generalize even when the property of the text to be parsed is more specific. This increases the theoretical relevance, but

[12] In classification it is necessary to define the topics or at least the number of topics. In scheduling it is necessary to define the dimensions to be measured. In corpus similarity, two texts will be similar in some way and different in others, since similarity can only be interpreted in the light of corpus homogeneity, that is, it is not appropriate, for example, to compare parliamentary speeches with technical manuals (KILGARRIFF; ROSE, 1998).

decreases the fidelity of the indicator (GRIMMER et al, 2018).

A common problem related to validation is called overfitting. When using the same documents to discover text properties, it is common to misinterpret the results. To solve this problem, one solution is to separate the training and test sets (split sample design). By splitting the samples and separating a training set for use in discovery (setting potential results) and a test set in estimation (analysis), the dependence between the discovery of the text properties to be analyzed and their causal effect is broken (GRIMMER et al, 2018). For Grimmer and Stewart (2013), the ideal validation procedure would be to divide the data into three subsets. The first would be the test subset, where the model would be fitted. The second would be the validation subset, hand coded and used to evaluate model performance. Finally, the model would be applied to classify the third subset.

Casas and Wilkerson (2017) also consider it helpful to train the algorithm on a set of texts already labeled before testing its accuracy on an unknown set. However, they believe that repeating this process over and over, using different sets for training and testing, and then aggregating the validation of results (N-fold cross-validation) is an even better approach. In addition to split sample design, another way to avoid overfitting is to use different algorithms to demonstrate if there are similar clusters (CASAS; WILKERSON, 2017, citing QUINN et al, 2010; GRIMMER; KING, 2011 and ROBERTS et al, 2014).

In addition to overfitting, there is also the issue of topic instability in unsupervised methods. To avoid this instability it is healthy to use results from different models of the same algorithm. Casas and Wilkerson (2017), for example, used 17 Dirichlet latent allocation (LDA) models – varying the number of topics between 10 and 90, in increments of five – to produce 850 topics (10 + 15 + 20… + 90). To determine which topics were consistent, they first calculated the cosine similarity for all topic pairs (resulting in 722,500 similarity scores) and then used the Spectral Clustering algorithm to group the 850 topics based on the similarity of cosine and check which topics remained constant (CASAS; WILKERSON, 2017).

Two other typical characteristics of parliamentary speeches are, as Lauderdale and Herzog (2016) demonstrate, sparsity and the selection of speakers. Considering that in real life only few representatives speak in certain debates, as a result, the Document Term Matrix – DTM – is sparse. Another point is party control and agenda control. To avoid this bias, Moreira (2016) used only Small Expedient speeches, and Schwartz (2018) used Small and Great Expedient speeches, because in these two moments, representatives are free to express themselves on any subject. Lauderdale and Herzog (2016) also found that in systems with strong incentives for personal and non-partisan voting, representatives speak more freely because parties recognize the need for recognition of representatives' names.

POSSIBLE APPLICATIONS AND RECOMMENDATIONS

There are several possible applications of automated content methods for Chamber of Deputies. Supervised methods, for example, may be better suited for saving labor – such as through automatic indexing or classification – and unsupervised methods for making discoveries and providing insight into these classifications (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017). However, they are also complementary methods, as supervised learning can be used to validate or generalize the findings provided by unsupervised methods (GRIMMER; STEWART, 2013).

In the case of supervised methods, it is evident that work can be saved by using automatic or at least semi-automatic indexing – if the server confirms the term suggested by the algorithm. Among the various indexing processes performed by the Chamber of Deputies, for example, are the indexing of parliamentary speeches (performed by the Department of Shorthand, Revision and Writing - DETAQ), of propositions (carried out by the Center for Documentation and Information - CEDI) and journalistic articles (carried out by the Secretariat of Social Communication - SECOM). In addition to indexing, other processes – such as joining and distributing propositions to committees (carried out by the General Secretariat of the Bureau – SGM) – could also gain much from the use of supervised methods. In addition to saving labor, automation would also bring greater uniformity to the indexing of the Chamber of Deputies and, consequently, facilitate the search and retrieval of information.

Therefore, it is recommended to use a supervised method – such as Naive Bayes – to implement the automation of indexing processes; as well as refining the classifier through a codebook, such as the Chamber Thesaurus base, and validating test results through manual coding iterations. Still, it is healthy to keep the human component in the work process to review the results presented by the machine. To avoid overfitting, it is recommended to use split sample design[13].

Another application of automated analysis for Chamber of Deputies is the provision of various strategic information. How important is it to know which topics were most debated by the Chamber of Deputies during the legislature? What is the position of representatives and parties on each of these topics? Is this a legislature mostly right or left? Liberal or conservative? Is there an ideological identification in the parties? Do representatives vote according to this identification? Is there consistency in what representatives say and how they vote? Is there party cohesion in parliamentary speeches? Is there alignment between the grassroots and opposition

[13] It is also possible to use the STM unsupervised method to discover new words and phrases that are frequently used by representatives, but not in indexing. This may be the case with neologisms and expressions such as "gender ideology", "school without party", "pixuleco", etc.

speeches? Who influences the legislative process and the speeches of representatives most? The executive branch? The parties? The electoral base? The interest groups?

The amount of information that can be extracted from parliamentary speeches – especially given the combination of these speeches with the products and outcomes of legislative activity – seems endless. For didactic purposes, it is possible to divide this information into three sets: (1) subjects or topics of debate, (2) positioning on a political or ideological scale, and (3) influence on the legislative process. How the use of words in political debate can vary according to the topic discussed, the classification of topics also serves as a dimension for analyzing political positioning and influence.

Thus, classification by topic emerges as Chamber of Deputies' first challenge in providing quality strategic information through automated content analysis. Among the various possible approaches, it is suggested to use supervised and unsupervised methods as complementary, using one algorithm to validate the results of the other. In this sense, there are two possible ways: (1) to use the STM unsupervised method for classification and to validate the results with the Naive Bayes/Decision Tree supervised method; or (2) use Naive Bayes / Decision Tree for classification and validate results with STM[14]. In both ways it is possible to use the binomial / trinomial result to reject the null hypothesis regarding the use of the two algorithms, as well as to validate the results with facts, events and results of legislative activity.

In addition, we recommend using split sample design for Naive Bayes, as well as refining the classifier and validating the training set through manual coding iterations. In the case of STM, it is recommended to use – following the teachings of Casas and Wilkerson (2017) and in view of the 32 themes used by the Chamber of Deputies in its leaderboard[15] – at least six different models with increments of eight topics each (8, 16, 24, 32, 40, 48). Thus, it is possible to solve the topic instability problem by grouping all topics resulting from the six models and checking which topics remain consistent.

It is important to remember that automated methods count the frequency of words in speech, not considering the time of speech. Thus, one representative may speak the same word dozens of times in a five-minute speech, and another may speak only a few times in a major speech. Although the use of lexical fields to define topics mitigates this problem, it is advisable to measure the debate time by topic – considered the start and end period of a debate as a speech metadata and to use it as an indicator of most debated topic in the Chamber of Deputies in a given period. It is believed that it is possible to achieve these results using an approach in which the time (start

[14] STM is the most recommended unsupervised method for allowing the use of metadata as covariates, such as speech time. Another important feature of this method is the possibility of creating a lexical field, such as a network of words, around the theme. Naive Bayes / Decision Tree is a well-tested algorithm whose use is already dominated by some Chamber of Deputies' servers.

[15] As the 32 themes are very broad, the definition of subthemes is essential to understand the parliamentary speech and to facilitate the classification of sets of propositions.

and end period) of the speeches is used as a covariate in the STM method. In this same sense, it is possible to measure the time of debate invested in a subtheme, a set of propositions, or a public policy. Thus, speaking time can be seen as an indicator of power by measuring the dominance of parties and representatives in the Chamber of Deputies' work, as well as a performance indicator by measuring the total time invested in debating public policy.

However, the topics discussed also reflect the ideologies present in Parliament. As Bakhtin (1988) would say, parliamentary speech speech is also an ideological content, it refers to something outside itself by establishing interests and establishing levels of domination, transforming Parliament into an arena where the battles of various national interests are fought. In this sense, ideologies express themselves not necessarily talking differently about the same problems, but also talking about different issues (DIERMEIER et al., 2012). Thus, we seek to understand through statistical models how these ideological contents manifest themselves in linguistic forms and how they reflect and / or refract such contents.

Empirically, ideology allows us to predict the political position of a representative on one subject through his position on another unrelated subject (DIERMEIER et al., 2012). Traditionally, political scientists use roll-call votes to estimate the individual positions of representatives. However, voting is often only symbolic and is not nominally registered. There is also strong party control in the polls to avoid dissenting votes. In this sense, it is more accurate to estimate the diversity of positions of representatives through the set of their speeches (LAUDERDALE; HERZOG, 2016). Parliamentary speeches would thus be a "product" that demonstrates the electoral connection.

Given that ideologies can manifest through the choice of topics, phrases or words in particular contexts; It is first necessary to define what is meant by ideology. Pragmatically, according to Converse (1964) ideology is a set of beliefs that guides the positioning of the individual in various subjects (DIERMEIER et al., 2012). This is a similar view to what Cancian (2007) advocates for empirical research: ideology is described as the set of ideas, values or beliefs that guide the perception and behavior of individuals on various subjects (SCHWARTZ, 2018).

By defining ideology as a set of beliefs that guide parliamentary positioning, it is possible to examine whether the ideological positions of representatives expressed in their speeches determine their votes, given institutional constraints such as agenda and party control, Diermeier et al. (2012). If the vote is explained more by a pre-existing political ideology (as expressed in the speech) than by institutional factors[16], knowing the position of representatives for a set of issues is predictive of knowing their vision for other uncorrelated issues[17] (DIERMEIER et al.,

[16] Diermeier at al (2012) compares what US Senators have said and how they voted, and concludes that voting and debating are different but correlated expressions of the same ideological belief system.

[17] How it has achieved a 94% accuracy in the rank of extreme senators and 52% in moderate senators – plus recognition, through additional experiments, that moderate senators are attenuated versions of extreme senators rather than a distinct category – it is believed that there is, in fact, a distinct lexical field for conservative and liberal legislators. Despite the

2012).

However, given the need for representative reelection – especially in systems with strong incentives for personal and non-partisan voting – representatives need, in addition to taking positions on various topics, to advertise to their electoral base and gain credit. At this point Casas and Wilkerson (2017) agree with Mayhew (1974) and conclude: Not every parliamentary speech is ideological. It is no coincidence that the research results of Diermeier et al. (2012) demonstrated that both Republicans and Democrats spend most of their speeches thanking voters[18]. Thus, the need to separate non-ideological content from the corpus as much as possible is evident, removing, for example, procedural speeches, tributes and thanks.

It is good practice to look at variations in word usage by topic before analyzing variations by placement. It is also advisable to use specific dimensions in each debate, then summarize the variation in these specific dimensions of each debate using two general dimensions – such as left/right, conservative / liberal or government / opposition. This allows you to retrieve multidimensionality in preference estimates with topic tags. Another good practice is to measure variation by topic and by representative, then merge all the topics of the same representatives to regain multidimensionality (LAUDERDALE; HERZOG, 2016).

As these practices depend on a detailed definition of the two (or more) general dimensions used, it is recommended to create a study group, with the participation of Chamber of Deputies' consultants, to create a dictionary annotated by topic and by public policy with the definitions and lexical field of what would be left and right, as well as conservative and liberal, on each of these topics – t in view of the indexing of the Chamber of Deputies and previous work on the subject[19]. In terms of methods, all three sets analyzed can measure positioning successfully. There are experiments with SVM – which can be used in conjunction with roll-call votes for this purpose – with Wordshoal – which combines Wordfish with Bayesian factor analysis to estimate positions – and with X^2 – text similarity approach combined with Bayesian factor analysis for the same purpose. It is recommended to use different methods to validate each other, as recommended in the literature.

In addition to measuring positioning, estimating political positions, and predicting voting, it is also possible to analyze the coherence of a representative – observing the variation of his / her positioning over time and the difference between his / her speeches and his roll-call votes –, as well as the cohesion of parties, blocs and parliamentary fronts. Party cohesion can be measured by the agglomeration of representatives on the ideological scale or the similarity of speech. In this

many issues discussed in Congress, conservatives and liberals always talk about any of these topics in distinct and stable ways (DIERMEIER et al, 2012).

[18] The results indicated that Republicans are more likely to give thanks (17%), followed by defense (8%), education (7%), and family (6%) speeches. Democrats also give more thanks speeches (9%), followed by speeches about education (9%), health (7%) and family (5%).

[19] Like Noberto Bobbio' book "Direita e Esquerda (Right and Left)"of (1994) and the project voteview.com.

sense, Diermeier et al. (2012) found evidence that party affiliation and ideological classification are highly correlated in the US, calculating "*kappa"* agreement as a measure of consistency between senators' ideological labels and party affiliation (there was almost perfect agreement: *kappa* = 0.932). A special type of cohesion is the alignment of the government base, as Schwartz (2018) has shown. In this sense, it is also advisable to focus on the use of different methods and approaches to measure the degree of party cohesion, as this type of information is of great importance to Chamber of Deputies' Leadership.

In addition to topics of debate and positioning on the ideological spectrum, there is a third use for automated methods: identifying influences on documents and outcomes of the legislative process. In more recent studies, influence is conceptualized as control over outcomes, more specifically the difference between what the actors involved expressly want and the outcome[20] (PRITONI, 2014). It can be difficult to unravel what the actors involved really want, but it is possible to identify the actors who have influenced public policy by knowing the origin of the ideas that are expressed in the approved legislation (BURGESS et al, 2016). Thus, it becomes possible and relevant to identify the influence of the Executive Branch, parties, blocs and parliamentary fronts, as well as interest groups, in the legislative process, through the reuse of text.

In this sense, Burgess et al (2016) used the Smith-Waterman algorithm and the ElasticSearch engine to trace the origin of propositions. This same approach can be used to verify text reuse in parliamentary opinions and amendments. Hertel-Fernandez and Kashin (2015) used the combination of three methods – involving similarity of bigrams and trigrams, LDA and SVM – to identify the influence of interest groups in US state law. However, these approaches focus on legislative products – such as propositions, opinions, amendments and studies – and on the legislation itself[21]. The use of Burgess et al (2016) approach in documents such as propositions and amendments is recommended. In addition to providing valuable information on text reuse, the results of this application can also contribute to the process of joining propositions and appraising amendments.

[20] However, Pritoni (2014) states that the actors involved may overstate their initial claims to facilitate bargaining and achieve the best possible outcome for the moment, regardless of whether the intention is to change or maintain public policy. Thus, only interest groups that want to maintain public policy would actually show their real intention, not to change anything.

[21] Regarding parliamentary speeches, similar methods can be used to trace the origin of words, expressions and arguments in parliamentary speeches. Such an approach, for example, may serve to analyze the similarity between the speech of representatives and the speeches of members of interest groups.

BIBLIOGRAPHIC REFERENCES

BAKHTIN, Mikhail. Marxismo e filosofia da linguagem. São Paulo: Hucitec, 1995.

BLEI, D. Probabilistic Topic Models. **Communications of the ACM**, New York, NY, USA, v. 55, n. 4, p. 77-84, 2012.

BURGESS, M. et al. **The Legislative Influence Detector**: Finding Text Reuse in State Legislation. *KDD*, San Francisco, CA, USA, 2016.

DIERMEIER, D. et al. Language and ideology in Congress. **Br. J. Polit. Sci**. 42(1):31–55, 2012.

GRIMMER, J. We are all social scientists now: how big data, machine learning, and causal inference work together. **PS: Political Science & Politics**, Cambridge, UK, v. 48, n. 1, p. 80-83, 2015.

GRIMMER, J.; STEWART, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. **Political Analysis**, Oxford, v. 21, n. 3, p. 267-297, jan. 2013.

GRIMMER, J. et al. How to Make Causal Inferences using Texts. *Working paper*, 2018.

HERTEL-FERNANDEZ, A.; KASHIN, K. 2015. **Capturing business power across the states with text reuse**. Presented at Annu. Meet. Midwest Polit. Sci. Assoc., Chicago, IL, Apr. 16–19, 2015.

IZUMI, M. **Velhas questões, novos métodos**: posições, agenda, ideologia e dinheiro na política brasileira. 2017. 113 f. Tese (Doutorado em Ciência Política) – Universidade de São Paulo, São Paulo, 2017.

KILGARRIFF, A.; ROSE, T. Measures for corpus similarity and homogeneity. **Information Technology Research Institute Technical Report Series**, Brighton, p. 46-52, jul. 1998.

KLUVER, H. Measuring interest group influence using quantitative text analysis. **Eur. Union Polit.**, v. 10, n. 4, p.535– 49, 2009.

KLUVER, H. **Lobbying in the European Union**: Interest groups, lobbying coalitions and policy change, Oxford, Oxford University Press, 2013.

KOHAVI, R. **Scaling Up the Accuracy of Naive-Bayes Classifiers**: a Decision-Tree Hybrid. **Data Mining and Visualization Silicon Graphics**, Inc. Mountain View, CA, USA, 2011.

LAUDERDALE, B., CLARK, T. Scaling politically meaningful dimensions using texts and votes. **Am. J. Polit. Sci**., v. 58, n. 3, p.754–71, 2014.

LAUDERDALE, B. HERZOG, A. Measuring political positions from legislative speech. **Political Analysis**, Cambridge, UK, v. 24, n. 3, p. 374-394, 2016.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge: The MIT Press, 1999.

MAYHEW, D. Congress: **The Electoral Connection**. Yale University Press, 1974.

MOREIRA, D. **Com a palavra os nobres deputados**: frequência e ênfase temática dos discursos dos parlamentares brasileiros. 2016. 204 f. Tese (Doutorado em Ciência Política) – Universidade de São Paulo, SP, 2016.

MOREIRA, D.; IZUMI M. O Texto como Dado: Desafios e Oportunidades para as Ciências Sociais. **Revista Brasileira de Informação Bibliográfica em Ciências Sociais – BIB**. São Paulo, BR, n. 86, 2/2018.

POWER, T.; ZUCCO, C. Estimating ideology of Brazilian legislative parties, 1990-2005: a research communication. **Latin American Research Review**, Pittsburgh, v. 44, n. 1, p. 218-246, 2009.

PRITONI, A. How To Measure Interest Group Influence: Evidence From Italy. **ECPR Joint Sessions of Workshops**, Salamanca, Spain, 2014.

ROBERTS, M. et al. The structural topic model and applied social science. **Advances in neural information processing systems workshop on topic models**: computation, application, and evaluation. Cambridge, MA: Harvard University, 2013.

ROBERTS, M. et al. Computer-Assisted Text Analysis for Comparative Politics. **Political Analysis**, Cambridge, UK, v. 23, p. 254-277, 2015.

SCHWARTZ, F. **Análise do Discurso Parlamentar por Meio da Técnica do Processamento de Linguagem Natural**: Abordagem Estatística e Aprendizagem de Máquina. 2018. 76 f. Pesquisa (Pós-Doutorado em Tecnologia) – Universidade de Brasília, BSB, 2018.

CASAS, A.; WILKERSON, J. Large-scale computerized text analysis in political science: Opportunities and challenges. **Annual Review of Political Science**, Palo Alto, v. 20, p. 529-544, 2017.

WILKERSON, J. et al. Tracing the flow of policy ideas in legislatures: a text reuse approach. **Am. J. Polit. Sci**., v. 59, n. 4, p.943–56, 2015.

Article received: 2018-27-12
Article Resubmitted: 2019-20-02
Article accepted for publication: 2019-12-03