Political speech analysis based on quantitative parameters: messages to Chilean congress from the military leadership and beyond

Análisis cuantitativo de los discursos políticos: mensajes presidenciales al congreso chileno desde el mandato militar al presente

Francisco A. Calderón[1]* https://orcid.org/0000-0003-2477-5858

Recibido 09 de diciembre de 2024, aceptado 7 de julio de 2025
Received: December 09, 2024 *Accepted: July 7, 2025*

ABSTRACT

Here is a quantitative analysis of word distributions in written messages to Congress over the past fifty years, encompassing eight presidential periods in Chile and one military leadership period. We have evaluated word rankings and distributions based on their frequency of repetitions, as well as fourth-order cumulant or binder cumulant. We believe Inter-words spacing implies a structural order similar in fashion to DNA markers organizing the text through the work of the author, embedded within the language. We found evidence for changes in distributions of the most frequent words, which may be responsible for the departure from Zipf's law, previously reported in Spanish. This effort aims to restore texts and create a more comprehensive compilation of messages sent to the Chilean Congress, with the potential for practical applications in optimizing word Corpus based on intrinsic distributions and rankings.

Keywords: Zipf's Law, quantitative linguistics, inter-word frequency.

RESUMEN

Se presenta un análisis cuantitativo de las distribuciones de palabras en mensajes anuales al congreso durante más de cincuenta años, abarcando ocho períodos presidenciales en Chile y un período de liderazgo militar. Hemos evaluado los rankings de palabras y las distribuciones de su frecuencia de repetición o espaciamiento inter-palabra, así como el cumulante de cuarto orden o de Binder. Suponemos que el espaciamiento entre palabras implica un orden estructural similar a los marcadores de ADN que organizan el texto a través del trabajo del autor quien elige, pero aun así dentro del lenguaje que le permite. Encontramos evidencia de cambios en las distribuciones de las palabras más frecuentes que pueden ser responsables de la desviación de la ley de Zipf, previamente reportada para el idioma español. Este trabajo es una restauración de los textos y compilación más extensa de los recursos disponibles como mensajes al congreso y al pueblo chileno cada año desde 1973, y su aplicación tiene el potencial de optimizar el corpus de palabras, tomando distribuciones intrínsecas extraídas desde cantidades estadísticas más frecuentes basadas en el propio corpus según ranking.

Palabras clave: Ley de Zipf, lingüística cuantitativa, espaciamiento inter-palabra.

[1] Universidad Católica del Norte. Departamento de Física, Antofagasta, Chile. E-mail: fcalderon@ucn.cl
* Autor de correspondencia: fcalderon@ucn.cl

DOI: 10.64966/ingeniare.v33.15

INTRODUCTION

Written language, as the expression of human knowledge, can provide a natural laboratory for the understanding of information, set in bulk or through simple expressions; the information is carried, understood, and processed when both parties are linguistically competent. Hence, language may be regarded as the primary tool for communication in the present day, a direct link to the brain of millions of people who use it in their daily lives, from performing simple and dedicated tasks under instructions, preserving cultural legacy or saving humanity's most precious knowledge, they are all passed through language and most times procured in written text. Several studies provide insight into the statistical distributions of words and their natural occurrence. Yet, we cannot deny the simplicity and power of a single word, such as "freedom" or "revolution," which conveys much more information in very few letters. The question of how information is embedded in written text is the subject of extensive studies and embodies such terms such as culture, history, and belonging (i.e., a specific group of people).

Hitherto, we may not only understand words and their specific meanings, but we also knowingly dive into interpretation once an accumulated number of words requires it, enabling many points of view, say, for any available paragraph. Through statistical models [1], we can observe how arranged word frequencies performed with reasonably good fitting to common parameters, for instance, in the English language [2].

The historical correspondence of this idea was popularized mainly by George Zipf and his Law for the ranking order of words in the written English language [3] and the work of Jean-Baptiste Estoup [4]. Some attempts to characterize other languages, such as Spanish, have revealed differences when compared to the English language, both at the tail and start of the distribution [5]. Also, a full array of books from several Latin-American famed authors showed evidence of two playing roles underneath the extent of the works, an indication for a more complex dynamic which was further studied after analyzing the inter-word repetition frequencies, evidencing lognormal distributions for the top most frequent words in the ranked order of the Zipf distribution [6].

Here, we present a quantitative analysis of the underlying dynamics of word distributions in written messages to Congress for approximately six decades, expanding eight presidential periods in the Chilean government and one period of military leadership. Estimates of word rankings and variable distributions of inter-word repetition frequencies enabled us to observe the intricate structural order, using words as markers that organize the text similarly as those of markers that organize DNA, but possibly characterizing the author's work. Remarkably, we believe that the change in distributions from the most frequent words may be responsible for the observed departure from Zipf's law fitting parameter, as previously observed in Spanish; this may be a feature that can provide historical evolution in the specifics of size and total number of different words presented in the corpus.

QUANTITATIVE MEASURES IN ANNUAL MESSAGES TO CONGRESS

Data

For this study, a collection of congressional speeches has been obtained through the public presidential congress library [7], accounting for over fifty speeches of different presidents, including some remarkable historical figures, such as the first Chilean Female president, Michelle Bachelet, a military period running Augusto Pinochet, and five other periods by Patricio Alwyn, Eduardo Frei, Ricardo Lagos, Sebastian Piñera and Gabriel Boric. Additionally, the comeback to power in two cases –President Piñera and President Bachelet– adds more similarity, one might expect, to the data. We obtained annual speeches directly from the web congress database from 1973 to 2023. The data was curated for typographical errors but otherwise maintained as it was presented to Congress and the Chilean population through public speeches each year.

METHODOLOGY

Initially, text processing is controlled by using an algorithm to remove snippets of text that qualify as codification impurities, such as those based on non-equivalent ANSI or Unicode characters. An example list may include most of the following characters found in the core data:

“(){}[]+?*|^$1234567890,.:;=¿’’”º¡@!–_°-«»“”’/~ª%†
¬–•ç⍰½▶--…ˆˋ<>”@ ÁÉÍÓÚ ABCDEFGHIJKLM
NÑOPQRTS UVWXYZ© ªлЛп сщэ‘. ”

In this way, uppercase and punctuation are removed, as well as characters that do not perform well under the sorting algorithm. These do not represent data of interest, as we are more focused on the repetition of words and their consecutive repetitions, from which we count the number of words inside. We called these two quantities Zipf's law and inter-word spacing or frequency of repetition [6].

Zipf's law is largely based on the principle of minimizing effort, which should provide aid in the information received and save time by repeating words. Therefore, ranking their number of appearances, one finds an 80-20 rule, where just a few words make the most significant contribution to the text in terms of word size. However, the quality and quantity of the information are not certain. In this way, the beginning of the rank is filled with mono-syllables, as shown in Table 1, and the numbers of words that are different, repeating less than a couple of times, will be found at the tail of the Zipfian distribution.

Zipf's argument on the principle of minimum effort had laid down a path for a family of power laws expanding across many areas [8] that established a common, maybe universal way, to accommodate not only word distributions but many other quantities found to follow a power law behavior, sometimes also known as Pareto's Law (income distribution), Lotka's law (citation index distribution) and more. It is required that the total number of words in a text becomes larger than those found in any office document with two or three pages. Instead we are focused here on tens of pages containing thousands of words, where the chances of repetition are much higher. We formulate Zipf's law as the total number of words N_T, equation (2), to represent a system or book or several books from one author or author. Given this condition, we present the Zipf's law as follows:

$$f_s = \frac{A}{s^{\alpha}} \tag{1}$$

$$N_T = \sum_{s=1}^{N_r} f_s \tag{2}$$

Where A is a normalising value, exponent α is the Zipf parameter, f_s is the frequency of each word, and N_r is the last rank value for corpus going from

Table 1. Frequency of 25 words until the name of the country appeared, as we might expect the word 'Chile' should come out higher over a speech to Congress, given by the president addressing the nation.

Word	Rank	Frequency
de	1	52907
la	2	26040
y	3	23749
en	4	19651
el	5	18009
que	6	16502
a	7	14751
los	8	12107
las	9	8672
del	10	8606
para	11	7898
se	12	7402
un	13	6893
con	14	6826
por	15	6136
una	16	4631
más	17	3586
al	18	3480
es	19	3357
no	20	2741
como	21	2603
este	22	2527
lo	23	2491
su	24	2458
chile	25	2400

$s = 1..N_r$. The cumulative value $N_X = \sum_{s=1}^{X} f_s$ is the total number of words N_X up to the value X, and it follows that $N_T > N_r$. Here N_T is the total number of words, different from N_r been the total number of different words. This cumulative distribution, N_X is also known to follow a power law [8], thus making Zipf's law much easier to understand using a log-normal plot. In this way, the value f_s goes from the most frequent to the least frequent in ascending value for the rank s, Table 1 second and third column.

The frequency of repetition or inter-word space is also present in every text once we have taken the required token and counted the number of words remaining in between. Hence, the number f_s can be

expressed in another way, $f_s = \sum_{r=1}^{N} C_r$; equation (2) in [6]. We introduce the frequency of repetition or inter-word spacing, C_r, obtained from the same text, using a different approach. Let us consider a simple example: taking a word as a marker, we establish a series of patterns. Consider the text, "The$_1$ fool$_2$ doth think he is wise$_3$, but the$_1$ wise$_3$ man knows himself to be a fool$_2$." If we choose marker 1 (the word 'the'), then we count seven inter-words until the next marker 1. Same thing with markers 2 and 3, so that we can have for marker 1 the spacing $r = 7$ repeating only once and $r = 8$ if we stop at the end of the sentence. Consider a whole book you have then completed C_r with a proper distribution, as shown in Figure 2. At the change of the marker, a new distribution arises each time, giving birth to a whole new way to understand the ordering in which words are entangled together according to the chosen marker.

It was reported that since the frequency decays as a power law, the capacity to produce a probability distribution using inter-word spacing falls exponentially [6]. Therefore, this quality of data is largely reserved to understand how the very few top words (mono-syllables) in Table 1 aid the writer in the task of transmitting information in an orderly fashion. Aligning with the observation that Zipf's law does not seem to match the whole range of words, but mostly at the tail [5], the idea for a minimum value x_{min} at which the Zipf's distribution should start has been considered [9], but without any supporting evidence this only seemed and an ad-hoc solution. The effort to provide for a more established system [10] to describe the amount of information or ambiguity drawn by words is highly desirable, even more so nowadays, with the computational power backing Artificial Intelligence and Neural Networks, both of which depend largely on relational properties among words.

Moments and cumulants are calculated under the same principle explained in [6] and [8]. The moments are defined under the most common of sense as $\left(x^m\right) = \sum_{x_{\min}}^{N} x^m p(x)$, leading to the definition of cumulants, which are the standard measures associated with any probability distribution. The cumulant arrangement and order are commonly known by the value of m, which is found in any standard statistical physics book. Hence, the most common cumulants are the mean value, the variance, Skewness, and kurtosis. Mayor order cumulants are possible, but we focused here on using these metrics to observe the changes in the moving average of the fourth cumulant or Binder cumulant as defined in [11], Figure 3.

RESULTS

We carried out the linear fitting of each year's text using equation (1) accounting from the year 1973 to 2023, as shown in Figure 1. Additionally, the inter-word distribution is then calculated for

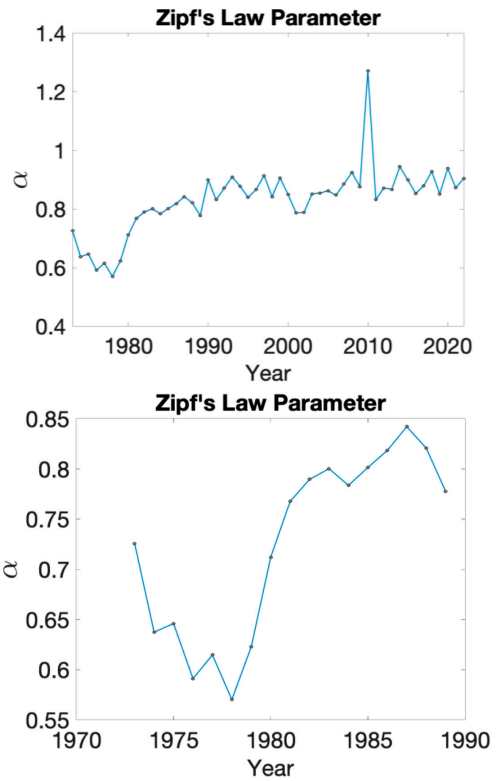

Figure 1. Linear fit of α parameter for each year, the power law α coefficient is plotted against years, or say different presidents. Since, α is commonly seen in literature for the English language [8] to have a value relatively close to 1.0, the deviation in the years up to 1981 are highly irregular. Top: The years 1973-2023 are presented, showing a slight variation from 1990 onwards. Down: Here, only the years 1973-1989 are presented, as they are regarded as the military period with large variations.

the first word in Table 1 and plotted along with its frequency distribution, against the inter-word spacing in Figure 2. This calculation was performed for more marker words but is not discussed further in this context. Thirdly, the Binder cumulant is shown in Figure 3 up to $s = 100$ since there are no other changes beyond this point where the curve found the transient. Lastly, a new Zipf's law was calculated using a grouping of words taken from specific values of $C_r = \{3, 10, 20\}$, as shown in Figure 4. Details are discussed below for all figures.

In Table 1, we tabulated the first twenty-five values of s according to equation (1) using the whole collection for better statistics. It is always remarkable how words with much more information or relevance do not scale up to higher values of frequency, such as the case for $s = \{25$: *'chile'*$\}$. We believe this is an indication of two clear regimes for the Zipf's law, as was already suggested by Cacho *et al.* [9]. Further evidence of this departure from Zipf's law in Spanish books was also shown by Calderón *et al.* [6] for several Latin-American authors.

In Figure 1, Zipf's parameter a is plotted against the corresponding year. We suspect this characteristic evolution for the value α has promising applications concerning the categorization of political speeches. First, we noticed the average value $\bar{\alpha} = 0.89$. Normally, α is close to unity for most texts written in English, and for Spanish, we have seen proximity to 1 as well; this indicates that Spanish may not strictly follow Zipf's law as proposed by George Zipf but has more information under the surface. Since the tail behavior was discussed in [5] and also an extended distribution was presented in the works of Mandelbrot [12], we prefer to focus here only on using α as a simple evolution parameter of the value changes from each year political speech, its corpus in extension and variety based on the fluctuations of the parameter α. The statistical dispersion $\sigma_\alpha = 0.081$ provides us with the limits [0.65,1.13] for $\bar{\alpha} \pm 3\sigma_\alpha$, providing now the upper and lower boundaries.

Interestingly, we observed an outlier point in 2010, representing a single event that exceeded the statistical limits for the period of return to democracy. Additionally, it is evident that from 1990 onwards, α has experienced very small fluctuations. Then, let us focus on the outlier event, which is passed $3\sigma_\alpha$. The event is related to a significant variation in the total number of different words N_r. With a little search, we discovered terms related to the

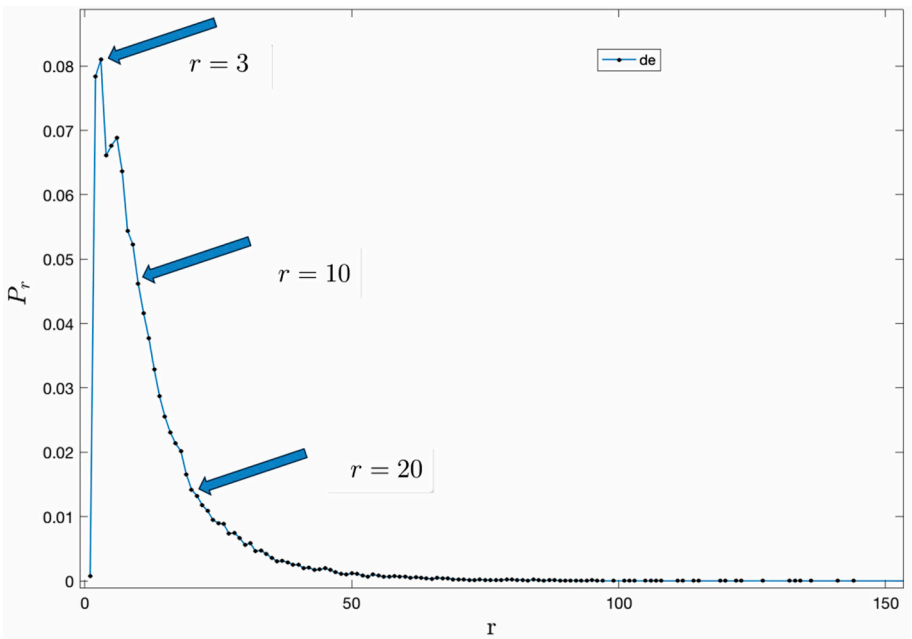

Figure 2. Probability distribution of repetition, or inter-word spacing.

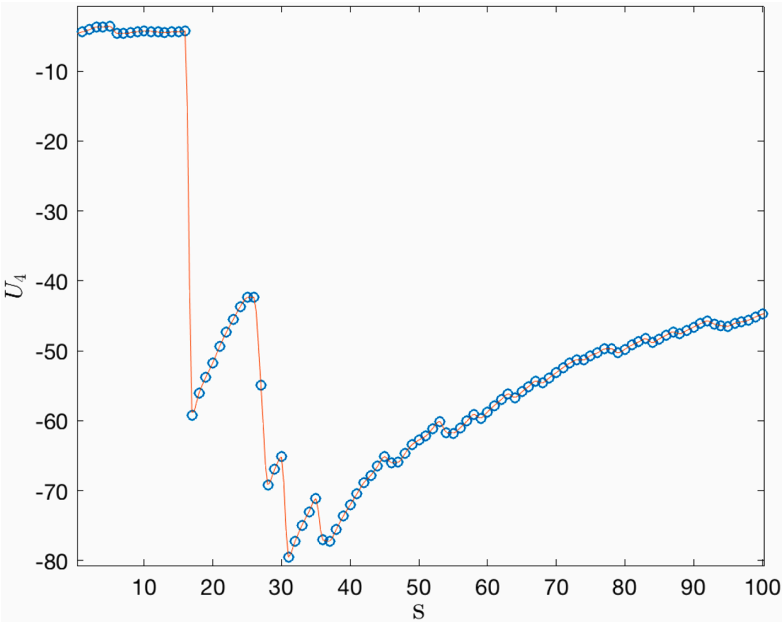

Figure 3. Moving average of Forth order Binder Cumulant calculated for the corpus.

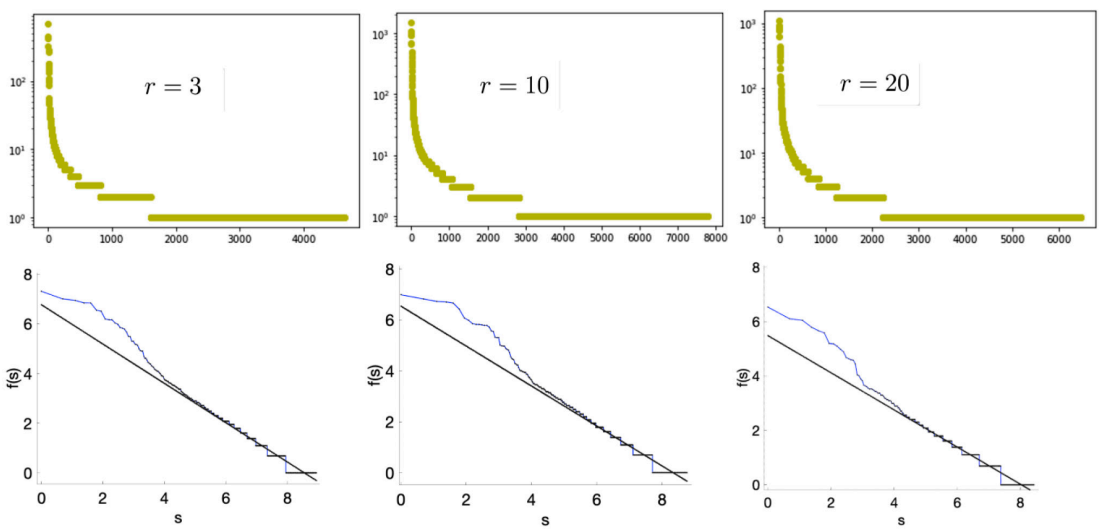

Figure 4. New frequency table, Zipf's law, calculated from certain values of $r = \{3, 10, 20\}$. This group of words from left to right does not intersect with each other.

devastating 8.8-magnitude earthquake that killed people and destroyed many areas in Chile. Not only the lives of the people were changed, but also how the president addressed the nation that year, which was observed in this significant deviation beyond $3\sigma_\alpha$.

Furthermore, we also observed that the lowest boundary was reached in 1978. However, this time, we can correlate to a decreasing then increasing value of the parameter, which implies that during the beginning of the military period, N_r was consistently lowered, with fewer different words every year up

to 1978. Then, the pattern reverses to a normal Zipf value by 1989. As one might easily find, Zipf's law was not intended to characterize anyone's words of preferences, but considering how each year the messages to Congress shall not necessarily bring a new corpus but maintain a similar one since the categories addressed by the president do not tend to change. Then, we can use the technique to identify the variations with external events, largely unexpected and impossible to predict, normally known as the 'Black Swan' effect.

In Figure 2, we show the frequency of repetition or inter-word spacing for the word 'de,' which is found to be the one with the highest frequency in all the corpus, making it the best possible probability distribution against r in increasing order. The logical situation here is for the most frequent inter-word spacing in this collection of text (1973-2023) to shed some light on preferred words with lower s. Assessing its ranking has the best statistical precision for $s = 1$, with decreasing precision as the number of markers decreases with rank [13] (higher s values). Notwithstanding, inter-word spacing r is a key value, and to comprehend Figure 2, it was discovered that information is accumulated as we take the cumulative distribution function for increasing values of r. Content in low r values can only be short phrases as simple as having two or three words (get out! or I am hungry) found between consecutive 'de'.

Nonetheless, most commonly, there is a tendency to write in the form of sentences, which typically contain about 6 or 7 words. Thus, we observe that in this type of political speeches the value $r \approx 3$ differs largely from the one observed in Latin-American Novels with $r \approx 6$ *or* 7 much more closely to our common sense. If we assume inter-words values of {3, 4, 5, 6 and 7} are ones comprising the most of the probability according to Figure 2, also another factor is that this repeats clearly all distributions ['de', 'la', 'y'].

However, the probability depicted in Figure 2 decays quickly, as a function of the value, nearly exponentially to zero beyond inter-word value, and so cases are approximately none-existent towards the end of this line. However, the probability depicted in figure 2 decays quickly, as a function of the value, nearly exponentially to zero beyond inter-word value, and so cases are approximately none-existent towards the end of this line. Thus, information seems to find most numbers of words around the central tendency. However, most sentences will have that many words in them, and as we move to the right, the long sentences should be less frequent, but also lose clarity as the increasing number of words obscures the meaning of very long sentences. Naturally, we are better at structuring paragraphs to make them easier to understand; this is something we are taught from an early age. More on this idea; as r becomes bigger, and we start having sections, then chapters, and finally a whole book, then the certainty over a group of words is again unclear, and we cannot consider any longer a set of words as a particular group in the book. In this way, we see that the few words and the entire book are representative of Figure 2, as it may imply that a corpus taken from specific ranges or values of r should carry much more information or meaning than initially expected.

In this way, we propose categorizing words of interest, those found in between markers, and using this technique to develop new groups of keywords or dictionaries, thereby optimizing rapid access to the most relevant information, as implied by this statistical approach. Since the preliminary results shown in Figure 2 are encouraging, new calculations are proposed for future works after testing these new ideas. Figure 4 for a quick approach.

Moreover, in Figure 2, we can see the probability distribution calculated from C_r using the marker 'de' plotted against an increasing value of inter-word spacing r. We group from here words found along the corpus, specifically those for $r = \{3, 10, 20\}$, as shown by arrows inside the Figure. We have calculated the fourth-order Binder cumulant as a running average, as shown in Figure 3; this remains unchanged for the first 16 points and then drastically falls, which is suspected to be highly dependent on the exponential nature of f_s as a decaying law. Nonetheless, the Binder cumulant, in its running average, seemed to reveal more evidence of a transition zone; therefore, we will continue to expand this idea. Nevertheless, for now, three of Zipf's laws are calculated with the groups previously mentioned, and they are shown in Figure 4. We can observe in the three plots a tendency towards power law behavior, and the fitting parameter *a is* correlated with the changes in r for the plots, shown from left to right.

CONCLUSION

We have selected several years of Political presidential messages to Congress and performed quantitative measures using key features, such as the historical evolution of the simple Zipf's law parameter, showing both clear departures from the ordinary value $\alpha = 1$ found within the limits of the English language. Secondly, we observed events with parameter values exceeding the upper limit, which can be linked to historically and socially significant events. In this case, we report that the 2010 Chilean earthquake had devastating consequences, and the parameter *a* deviated largely from the statistical fluctuation, reaching a value $\alpha = 1.28$ nearly as far as the $4\sigma_\alpha$ limit, evidence that a significant number of new words, not frequently used in traditional messages, were employed in such occasion by president Piñera. Also, over the other extreme, military leadership was found to have departed from the Zipf's average parameter, and it was slowly recovering this tendency by the end of the period. Fourth order cumulant, although not significantly show evidence, does question the argument of having an x_{min}, and favoured more for a two regimes model. Finally, the measured values for Zipf's law coefficients using only reduced corpus, has become itself a power law distribution, and its parameter correlates inversely with the value of inter-words spacing. Thus, there is evidence that quantitative techniques can shed some light on political discourses [14]-[15]. Hence the historical evolution of commonly made messages to Congress could, in principle, provide supporting evidence of changes in central tendencies and their intrinsic information, known as corpus N_T.

ACKNOWLEDGMENTS

We would like to thank "Núcleo de Investigación No. 7 UCN-VRIDT 076/2020, Núcleo de modelación y simulación científica (NMSC)" and "Núcleo de Investigación No. 2 - UCN-VRIDT 042/2020 - Sistemas Complejos en Ciencia e Ingeniería" for the scientific support. Computational work was partially supported by the supercomputing infrastructures of the NLHPC (ECM-02).

REFERENCES

[1] H. Baayen, "Statistical models for word frequency distributions: A linguistic evaluation," *Computers and the Humanities,* vol. 26, pp. 347-363, 1992, doi: 10.1007/BF00136980.

[2] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing,* Cambridge, MA, USA: MIT Press, 1999.

[3] G.K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA, USA: Harvard University Press, 1932.

[4] J.B. Estoup, *Gammes Stenographiques*, Paris, France: Institut Stenographique de France, 1916.

[5] M.A. Montemurro, "Beyond the Zipf-Mandelbrot law in quantitative linguistics," *Physica A: Statistical Mechanics and its Applications*, vol. 300, no. 3-4, pp. 567-578, 2001, doi: 10.1016/S0378-4371(01)00355-7.

[6] F.A. Calderón, S. Curilef, and M.L.L. de Guevara, "Probability distribution in a quantitative linguistic problem," *Brazilian Journal of Physics*, vol. 39, p. 500, 2009, doi: 10.1590/S0103-97332009000400028.

[7] Biblioteca del Congreso Nacional de Chile. "Diarios de Sesiones del Congreso Nacional: Mensajes Presidenciales ante el Congreso Pleno", bcn.cl. Accedido: 6 de junio de 2024. [En línea]. Disponible: https://www.bcn.cl/historiapolitica/corporaciones/cuentas_publicas/detalle?tipo=presidentes

[8] M.E.J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323-351, 2005, doi: 10.1080/00107510500052444.

[9] R.F.I. Cancho and R.V. Solé, "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited," *Journal of Quantitative Linguistics*, vol. 8, no. 3, pp. 165-173, 2001, doi: 10.1076/jqul.8.3.165.4101.

[10] R.F.I. Cancho and R.V. Solé, "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences,* vol. 100, no. 3, pp. 788-791, 2003, doi: 10.1073/pnas.0335980100.

[11] S.H. Tsai and S.R. Salinas, "Fourth-order cumulants to characterize the phase transitions of a spin-1 ising model," *Brazilian Journal of Physics*, *vol.* 28, p. 1, 1998, doi: 10.1590/S0103-97331998000100008.

[12] B.B. Mandelbrot, *The fractal geometry of nature*, vol. 1, New York, USA: WH freeman, 1982.

[13] G. Iñiguez, C. Pineda, C. Gershenson, and A.L. Barabási, "Dynamics of ranking," *Nature Communications*, vol. 13, Art.no. 1646, 2022, doi: 10.1038/s41467-022-29256-x.

[14] E.C. Tucker, C.J. Capps, and L. Shamir, "A data science approach to 138 years of congressional speeches," *Heliyon*, vol. 6, no. 8, Art. no. e04417, 2020, doi: 10.1016/j.heliyon.2020.e04417.

[15] D.M. Mayaffre and C. Poudat, "*Approaches to complexity in European political discourse,*" in *Speaking of Europe*, K. Fløttum, Ed., Amsterdam, Netherlands: John Benjamins, 2013, pp. 65-83.