

Words ranking and Hirsch index for identifying the core of the hapaxes in political texts

Valerio Ficcadenti^{1,*} Roy Cerqueti^{1,2} Marcel Ausloos^{3,4}
Gurjeet Dhese¹

¹ School of Business
London South Bank University
London, SE1 0AA, UK
`ficcadv2@lsbu.ac.uk; cerquetr@lsbu.ac.uk; dhesig@lsbu.ac.uk`

² Department of Social and Economic Sciences
Sapienza University of Rome
Rome, I-00185, Italy
`roy.cerqueti@uniroma1.it`

³ School of Business
University of Leicester
University Road. Leicester, LE1 7RH, UK
`ma683@le.ac.uk`

⁴ Department of Statistics and Econometrics
Bucharest University of Economic Studies
Bucharest, Romania

Abstract

This paper deals with a quantitative analysis of the content of official political speeches. We study a set of about one thousand talks pronounced by the US Presidents, ranging from Washington to Trump. In particular, we search for the relevance of the rare words, i.e. those said only once in each speech – the so-called *hapaxes*. We implement a rank-size procedure of Zipf-Mandelbrot type for discussing the hapaxes' frequencies regularity over the overall set of speeches. Starting from the obtained rank-size law, we define and detect the *core of the hapaxes* set by means of a procedure based on an Hirsch index variant. We discuss the resulting list of words in the light of the overall US Presidents' speeches. We further show that this core of hapaxes itself can be well fitted through a Zipf-Mandelbrot law and that contains elements

*Corresponding author.

producing deviations at the low ranks between scatter plots and fitted curve – the so-called *king* and *vice-roy effect*. Some socio-political insights are derived from the obtained findings about the US Presidents messages.

Keywords: Text analysis; H-index; Rank-size law; Hapaxes; US Presidents speeches.

1 Introduction

Scientific debate has recently grown on text analysis and data mining because of the relevance of the information taken from texts and for the need of a systematic quantitative analysis of them. For example, it is worth mentioning [38], where the authors study the regularities of words occurred in blogs and [31], where the authors propose a model for assessing borrowers' defaults on loans by analyzing texts on the available descriptions of such loans. In [13] the authors pay peculiar attention to the exploration of the financial texts for their relevant informative content. In the same context, in [78] there is a discussion of the determinants of the crowdfunding successes by following a text analysis approach.

Nowadays, politicians use social networks to inform their voters, therefore they calibrate the messages on the bases of objectives to be addressed (see [77] for the case of Seoul mayoral election). The official speeches of the US Presidents are, of course, carefully written. Each single locution or term is evaluated, in order to guess what the impact will be on the audience and in the entire socio-economic environment.

This paper begins with the above premise and applies it to the analysis of some relevant aspects of a large set of Presidents' speeches. In doing so, we are in line with the studies of the communication of US Presidents (and also candidates to presidency or Presidents' media importance [58, 72]) and its socio-economic relevance (see e.g. [15, 22]).

Our target is to assess the presence of regularities in the frequencies of the hapaxes and explore the existence of a qualified set of words pronounced only once in a large number of speeches. Such regularities are able to outline a scheme for supporting decisions in communication task. It is important to point out that the methodology here designed can be used in any type of corpora whose hapaxes follow a rank-size power law behaviour. With this aim, we study the collection of the hapaxes in each speech.

But why are hapaxes so relevant in the official context of the US Presidents' speeches, and why is it worthy to explore them? Hapaxes might represent subliminal-like messages that the US Presidents deliver to the audience. Furthermore, hapaxes play a fundamental role in identifying the richness of a speech. Thus, the use of hapaxes represents an instrument for informing the audience that

the President has a high level of culture and education, hence giving him credit with electors and relevant institutional figures. Under this perspective, it is important to mention one of the most commonly used morphological productivity measure, i.e. (see [10]):

$$P = \frac{\text{number of hapaxes in a text}}{\text{length of the text}}.$$

It is also important to stress that the most part of the hapaxes is associated to some peculiar characteristics, which lead to crucial research questions and conjectures (see the long list of future research outlined in the last section).

The dataset here considered comes from the rough data contained in the Miller Center, which is a research institute affiliated to the University of Virginia (see <http://millercenter.org>). A set of about 1000 US Presidents' speeches has been downloaded from such a website, from the *Inaugural address* of George Washington (1789) to the Donald Trump's speech *Address to Joint Session of Congress* (2017). Presidents' contributions like [55] are excluded; indeed they do not appear into the Miller Center database and they are not framed as speeches, but as scientific papers.

The treatment phase of the speeches – whose details will be presented in Section 3 – has been attained through data mining techniques (for a survey on text mining, see e.g. [20, 46]).

The study proceeds in three sequential directions (steps).

Firstly, a rank-size relation has been assessed over the set of the hapaxes, where *size* is measured through the frequency of the words in the entire set of speeches. For an excellent review of the empirical settings where power laws is a valid device for representing related phenomena and some theoretical explanations of such a way to fit data, we refer to [54]. In agreement with other linguistic studies [49, 61, 1, 2, 8, 65], we have tested the validity of the Zipf-Mandelbrot law in properly fitting the data by implementing a best-fit optimization procedure [79, 80, 43]. In this preliminary phase, we have found statistical compliance of the considered dataset with Zipf-Mandelbrot law, even in presence of (quite) negligible deviations at low ranks (see the fourth step below for a comment on this).

In the second step, we have used the obtained calibrated curve to identify the core of the hapaxes by using the indicator proposed in [5], in the science measurement context of the scientists' coauthors. Such an indicator is a replication of the *H*-index – where *H* stands for Hirsch, who invented it in [25] – used to evaluate scientific research (see [24] for a more detailed description and [69] for a comparison with different variants). In this context, the core of the hapaxes is the set with cardinality $\bar{H} \in \mathbb{N}$ which contains the maximum number of hapaxes whose frequency is at least \bar{H} . In this respect, the ratio between the area of the core and the one of the entire set of hapaxes – computed with respect to the best-fit curve of the rank-size law – is a percentage measure of the most relevant hapaxes in the overall history of the US Presidents' speeches.

The third step consists in the exploration of the core and of its properties. We here show that the core is a set whose hapaxes have ranked frequencies again satisfying a Zipf-Mandelbrot law. Furthermore, as already pre-announced above, in the present rank-size analysis of the overall set of the hapaxes, we have found small deviations at low ranks. This means that the best fit curve does not represent "perfectly" the scatter plot of the low-ranked hapaxes. The reason for this stands in the outlier-type behaviour of a group of hapaxes which are contained in several speeches. We here guess that such outliers are the hapaxes in the core and redo the best fit procedure by removing the core from the overall original sample. Results confirm the improvement of the fit. According to [37], the token at rank equal to 1 is the so-called *king* whilst the others are the *vice-roys*, and in this case there is a *king plus vice-roy effect*. For a further example of this effect, refer to [12]. An interpretation of such an effect will be also presented in the Section devoted to discuss the results. To sum up, the contributions of this paper relate to the well known rank-size behaviors of words frequencies in textual data; in particular, we address the peculiar case of the hapaxes distribution in corpora. We propose an innovative method to determine relevant hapaxes; the approaches used and the findings have a comparable impact to that of [50], where the authors propose an innovative approach to analyze opinions in social network and to determine decision making processes on the basis of users' sentiments.

We point out the presence of a link between this paper and [21]. Indeed, analogously to the quoted paper, we here move from the discourses retrieved from the Miller Center website, and opportunely treat them to extract individual constitutive words with the related frequencies for each of them. Differently, we here focus only on the words said only once in any discourse. In so doing, we are radically different from [21] in three main directions: firstly, about the data, the object of the analysis of the present paper is a subset of the source database used in the quoted paper. Here we analyse just the words with frequency one in each speech, while in [21] all the words are considered. Consequently, the data processing phase is a refinement of the one employed for collecting and treating all the pronounced words; secondly, the scientific ground of [21] lies in the aim of understanding and comparing the regularities of the structures of the individual speeches in terms of words frequencies, while the ground of the present report lies in understanding the way in which the same words – i.e., those used only once in each speech – have been historically employed by the US Presidents; thirdly, sometimes the Presidents want to communicate messages with a certain degree of discretion, targeting a specific subset of the audience. In this respect, the hapaxes analysis maps the unique reference to less central topics treated in a speech. Their investigation, especially on such a dataset, gives hints about the informative content and the target of the presidential communications. In so doing, the results provides a wide number of research questions, to be addressed in future studies.

The rest of the paper is organized as follows. Section 2 contains a wide discussion on the reference literature. In Section 3 there is a presentation of the procedure employed to collect the data. Section 4 is devoted to the illustration of the methodology used for the analysis. Section 5 presents the results and related comments. The last section offers some conclusive remarks; moreover, it contains also a specific focus on future research themes.

2 Literature review

This section contains a brief review of the literature, to support the research and scientifically motivate the worthiness of the proposed study. We start with a discussion of the hapaxes; then, we present a critical view of the employed methodology.

The exploration of the hapaxes goes much further than usual text content or structure analyses; hapaxes have a special meaning (see e.g. [6, 57, 19, 4, 16, 45]). Some remarkable examples are worth mentioning.

In the overall work of Giacomo Leopardi, the word *ultrafilosofia* has been used only once for the contextualization of the philosophical system of the author. However, the authoritative Encyclopedia Treccani refers to *ultrafilosofia* to describe Leopardi's thought. In the related entry, *ultrafilosofia* is no longer a hapax, but appears 9 times [62].

Mnemosynus is a hapax for the Latin language. In fact, it appears in the entire collection of available writings in Latin only once, in Catullo's Carmina. This term points to the mythological figure of the goddess of memory. Such a hapax has been not neglected in subsequent modifications and contaminations of the linguistic evolution, and *mnemonic* comes evidently from *Mnemosynus*. It is important to mention also the relevance of the hapaxes in the holy books of Bible and Quoran [30, 73], which contain speeches attributed to one or several authors.

Thus, it is not unexpected that some authors have dealt with the analysis of the hapaxes.

In [32], the authors state that the presence of hapaxes in a text can be used to determine the language productivity of terms, so the language inflection (see also [10] for related material). Therefore, studying the hapaxes of corpus from a common source along the years allows to capture neologisms (e.g. [9]). Consequently, it is useful to interpret changes into a community of people speaking a common language as presented in [41]. The study of the hapaxes across different documents is important in authorship attribution as well (see [28], where a wide description of the field is reported). For example, in [27], the author states that it is possible to test the tendency of an author to choose between a word used previously or utilizing a new word instead. In [70] is outlined the importance of the hapaxes within a corpus to extract information about the writers

stylistic changes.

To highlight the potentiality of the hapaxes analysis in spreading political messages and in supporting decisions about messages' structures, it is worthy to mention the different Figures of Speeches (for a review of them, see [71]). Indeed, there are specific figures like the *climax* used to increase the importance of a concept by intensifying the usage of certain words. On the other hand, by negation, if the message has to be hidden rather than stated, mentioning a word just once is enough to relay the communication. The analysis of hapaxes has never been employed in this respect but the hapaxes legomena are often involved in political communications analysis, e.g. see [67] for the specific case of US President candidates John McCain and Barack Obama.

In general, some political speeches have hidden embedded information about the future, see for example 'Address at Moscow State University' stated in May 31, 1988, by President Ronald Reagan, less than one year before the fall of the Berlin Wall, where the words "freedom" is repeated more than 20 times. Is it significant or related to the fall of the Berlin Wall occurred shortly thereafter? The US President represents one of the large powers in the world, so it is easy to guess that they have a large competitive advantage in gaining information with respect to their audience (President Nixon is a clamorous and ironic example of that!). The US Presidents have classified information about potential wars and treats for the country, so they might aim at preparing the public opinion for potential conflict by evoking specific sentiments. For example this could be the case of US before the World War II, when President Roosevelt started to consider an involvement of US into the Britain - French and German conflict (see the Fireside Chat 15 and 16 of President Roosevelt, the former titled 'On National Defense' and the latter 'On the Arsenal of Democracy').

Therefore, in the very special context of political communication, we notice that the US Presidents' speeches are written in a so precise and careful way that the presence of a word pronounced only once cannot be seen as accidental. However, the occurrence of a hapax in one speech does not tell credible stories *per se*. Despite this evidence, the presence of regularities in the selection of a given hapax over a wide number of speeches points out to a common behaviour of the Presidents, and merits attention. This is the scientific ground of our study.

We now provide a critical discussion of the methodological devices used in the paper, along with the most relevant literature contributions.

The employment of regression techniques of rank-size type in the text analysis finds also support in the literature. For the special case of the exploration of hapaxes within a rank-size framework, we refer to [48, 51], where the authors employ Zipf's law and point out that the relevant analyses of hapaxes, helps in the exploration of parallel documents.

In this respect, it is important to note that rank-size analysis allows to derive a panoramic view

of a unified system generated by granular data – the frequencies of the hapaxes.

We also point out that the definition of the core through the Hirsch H -index exploits the meaningfulness of such an index. In particular, the H -index is able to synthesize the overall number of hapaxes and the frequencies with which they appear in the set of the speeches in a unique entity. To fully understand this point we refer to the familiar use of H in the research evaluation context, where H gives a clear idea on the overall productivity of a scientist and on the impact – in terms of citations – of her/his production on the scientific community, i.e. the so-called "core of publications". In this line, one can extend the Hirsch-index idea to other cases, like to define the "core of coauthors" of a researcher (see [5, 6]). Generalizing the idea, we define the core of hapaxes in the investigated texts. Let us briefly recall the logical aspects.

Indeed, it was found out that a Zipf-like law

$$J \propto 1/r, \tag{1}$$

exists, between the number (J) of joint publications (NJP) of a scientist, called for short "principal investigator" (PI) with her/his coauthor(s) (CAs); $r = 1, \dots$ is an integer allowing some hierarchical ranking of the CAs; $r = 1$ being the most prolific co-author of the PI. Yet, it was observed that a hyperbolic (scaling) law is more appropriate, i.e.,

$$J = J_0/r^\alpha, \tag{2}$$

with $\alpha \neq 1$, usually such that $\alpha \leq 1$, and often decreases with the number of CAs or with the number of joint publications, e.g. when the number of CAs and when J are "not large". J_0 is a fit parameter, i.e. there is no meaning to $r = 0$.

We can follow such a line of thought for the hapaxes, being ranked, and noticing those below a given threshold. As the H -index [25, 26, 64] "defines" the *core of papers of an author* from the relationship between the number of citations n_c and the corresponding rank r of a paper, through a trivial threshold, i.e. if $n_c \geq r_c$, then $r_c \equiv h$, thus one is allowed to define the *core of hapaxes* through a threshold [5, 6], called the m_a -index, for short,

$$m_a \equiv r, \quad \text{as long as } r \leq J. \tag{3}$$

Technically, one could thus measure the relevant strength of a hapax, whence measure some impact of such a word on speech intention, as in research collaborations [39]. This is exactly the methodological perspective adopted in the current paper.

3 Data

This section is dedicated to describing the data collection process. It is very close to the one presented in [21, 14], hence we report here just the phases implemented to get the hapaxes and to improve the original procedures. The key process' steps are summarized in Figure 1.

As a premise, a methodological remark is in order. Of course, the adopted strategy for analyzing the speeches includes also personal visual inspection. However, the general procedure for the pre-processing and web scraping phases are stepwise listed. In this way, it is possible to replicate the study.

Data have been retrieved from the Miller Center, which is a research center affiliated to the University of Virginia (see <http://millercenter.org>). The Miller Center is still active and adds continuously speeches. Thus, the codification of the problems fosters future updates of the analysis of the hapaxes.

We point out that a conservative principle drives our actions, so that the original texts are modified as little as possible. In so doing, we go in the direction of maintaining the highest level of similarity of the final dataset with the original speeches. At the same time, all the errors into the transcripts – including those of minor nature – are taken into full consideration. This required additional efforts with respect to [14], in fact, we have checked the text by using the “*hunspell*” R package [56], with the English dictionary. We extract and correct all the spelling errors within each talk. Nevertheless, there is still a list of 7716 exceptional tokens that are not found in the English dictionary. They are considered as potential typos that have to be investigated one by one. Many of them are not exactly errors, but they are exceptions invoked by the speakers for satisfying specific rhetoric needs or past ways of speaking not comprised into the US Hunspell English dictionary. Some examples are given by the usage of peculiar non-English personal names like “Bernardino”, terms from Spanish or French like: “intendencia” or “arrete” and terms that were differently spelt in the past like: “regrassing” or “tofore”. Even if these types of exceptions enter into the list of potential typos, they cannot be considered errors to be modified because we assume that the Presidents have pronounced them into a specific context that required such uses. Therefore, to make the distinction between the exceptions just described and the flaws that have to be modified, we have looked at all the lists of potential typos into the *Cambridge Dictionary*, the *Oxford Dictionary* or *Wiktionary*. In this way, it is possible to identify ancient English uses no longer in vogue, foreign words or common language flaws in accord to the proper linguistics uses. When there are not any straight suggestions, one has to extract the entire phrase that contains the ambiguous terms from the respective transcript (the whole statements are easily captured by looking for them into the corpora). Then, thanks to the exact search of *Google*, it is possible to check if the ambiguous words

are reported into other speeches' transcripts sources by examining the results of the research. If there are other references for the same phrases with the terms corrected, we adopt the most logical usage by interpreting the meanings of the findings¹. The potential flaws with the highest degree of uncertainty are adjusted or not on the bases of the majority criterion. It means that one uses Google's exact search and then adopts the most common phrases within the first ten results hereby found. Thanks to this correction process, 3851 improvements are applied. Consequently, the remaining locutions cannot be considered wrong. This step is the only one requiring human judgment; therefore, it is the only reason to define the nature of the procedure as *semi-automatic*.

The whole resulting list of words pronounced only once per speech is made of 31074 tokens, with frequencies that range in [1, 250]. It means that there is a term used just once in 250 speeches and another that appear one single time only. The principal statistical indicators can be found in Table 1, column (a), which allows having a view of the frequencies' collection properties. Figure 2 shows the percentage of hapaxes per speech along with the speaker's affiliation (this last information is there just to enrich the information carried out by the figure).

Despite the applied correction process, some minor typos are still reported into the hapaxes dataset. They do not exceed 2%. This level of error is expected for a dataset made by speeches transcriptions and it is unavoidable also for linguistic reasons. Indeed, we are not able to resolve some linguistic ambiguity of certain terms, whence inducing some inaccurate error bar. However, we used computer-based algorithms and human readings to reach an expectedly very low set of possible forgotten cases of misspellings. For example, the word "unmunitoned" which appears in "Address at the Celebration of the 150th Anniversary of George Washington Taking Command of the Continental Army, Cambridge, Mass" stated by Calvin Coolidge on July 3rd, 1925, can be a misspelt word or not. We prefer to consider it as a typo contributing to the stated 2% since we are not sure of its meaning; a linguistic investigation – unmonitored, at the moment – would be appropriate for cases like this. Furthermore, even if one decides to go through each residual terms, there remains the possibility of leaving errors due to natural human predisposition to commit an error in a manual control – operational risk. The idea of using a coded routine for spell checking the ambiguous words is a side scientific product of the research; this is an improvement to [21].

¹Examples of this type are given by the typos "questionin", "lawon" and "adispute", that come from the Inaugural Address stated by Rutherford B. Hayes, March 5, 1877. The bugs are in the following phrase: "*The fact that two great political parties have in this way settled adispute in regard to which good men differ as to the facts and the lawno less than as to the proper course to be pursued in solving the questionin controversy is an occasion for general rejoicing.*" One can intuitively guess that the corrections of the wrong terms are "question in", "law on" and "a dispute" but for acting systematically and for being coherent with the method adopted, the exact search is run. In the context of the example, the research returns many other sources where the words' correct forms are adopted as expected.

| Statistical indicator | Whole corpus (a) | Core hapaxes (b) |
|---------------------------------|------------------|------------------|
| N. Words | 31074 | 182 |
| Mean (μ) | 16.3850 | 199.6484 |
| Variance (σ^2) | 1034.2965 | 183.279 |
| Standard deviation (σ) | 32.1605 | 13.5381 |
| Skewness | 3.2451 | 1.1188 |
| Kurtosis | 11.5989 | 1.463165 |
| Median (m) | 3 | 197 |
| Max | 250 | 250 |
| Min | 1 | 183 |
| RMS | 36.0934 | 199.644 |
| Standard Error | 0.1824 | 1.0035 |
| μ/σ | 0.5095 | 14.7472 |
| $3(\mu - m)/\sigma$ | 1.2486 | 0.5869 |

Table 1: Main statistical indicators associated to (a) the whole list of hapaxes in the dataset and (b) the set of the hapaxes belonging to the core.

In this way, we make a procedure that defines what can be considered as typo into the context of this framework. Anyway, a visual inspection of the remaining terms allows concluding that the majority falls into the terms that occur just once into the hapaxes’ list. Therefore, this reinforces the idea that the residual cases lead to a negligible effect on the analysis object of the present study.

Yet, as we will show in the section devoted to the methodology, the procedure to determine the analyze the hapaxes and identify their “core” combines the H-index (see [25]) with the Zipf rank-size law. In particular, the core consists of a group of hapaxes which are the most frequent ones in the US Presidents’ speeches dataset. The definition of the core guarantees that the possibly existing typos are excluded from it, and can be found just in the tail of the list of the hapaxes (sorted by frequency). Indeed such tokens, being typos, have very low chances of occurring more than once (namely in more than one speech), and appear only in the speech containing them.

Besides, even if the presence of typos in the tail of the hapaxes list could affect the estimation of the parameters of the Zipf-Mandelbrot law, a so small percentage of 2% is statistically considered not to lead drastic changes which could be harmful to the analysis.

Before moving to describe the methodological devices used for the analysis, a remark on the worthiness of the dataset is needed.

The corpus we have used in this study is the same corpus that has been employed also in [14, 21]. To the best of our knowledge, it is one of the most complete that it is possible to find in the literature. To support this point, we mention studies in which the authors have analysed institutions' communications:

- In [67], the authors have used 245 US Presidents' speeches.
- [66] is an outstanding study of the State of the Union discourses in US. The set of analyzed speeches is a sensible subset of our corpus.
- In [40] the analysed data consists of the corpus of United States' presidential Inaugural Addresses from 1789–2009. Also in this case, the dataset is a sensible subset of our corpus.
- The contribution [68] focuses only on the Ronald Reagan's discourses.
- In [34], the authors have used about 150 Central Banks communications. They analyzed different monetary institutions, but for each of them they have used about 150 communications.

It is also important to point out that it is frequent to find researches whose target or scientific analysis ground is given by a restricted number of communications delivered by institutions and political subjects. Unfortunately, the absence of digitalization of old texts sometimes prevents the employment of a complete corpus mainly for difficulties in having access to the original texts or for the lack of scanning systems and precise enough tools like the latest OCR (Optical Character Recognition) technologies. In the case treated in the present paper, we are free from this problem, since almost all the speeches delivered during the US history are available and accessible on a website. Furthermore, the procedure designed in this paper is particularly appropriate to explore corpora made by a non-enormous amount of texts.

4 Methodology

The hapaxes of the individual speeches have been merged together in a unified list. Each word is associated to an integer, which assigns to it the number of speeches in which such a word is a hapax. We briefly call *frequency* this number, so that the resulting list is composed by a series of words with associated frequencies.

The resulting list of hapaxes contains 31074 words. The maximum frequency is realized by the word *sense*, which appears 250 times as a hapax in a President's speech. Moreover, there is a list of 10088 tokens which are hapaxes only in one speech (thus, having unitary frequency).

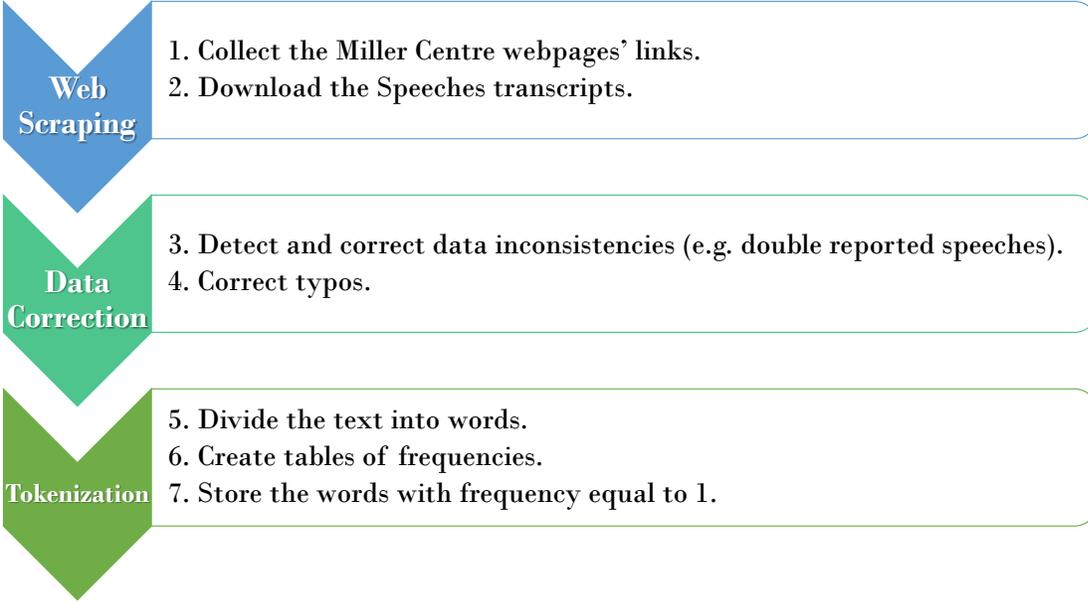


Figure 1: Summary of the necessary phases to collect and store the data used in this paper.

For the details of the construction of the dataset, see the Section 3 of this paper.

Hapax words are ranked in decreasing order, according to their frequencies. In this respect, the "size" of a word is its frequency. In the rank-size analysis, we will denote size and rank by s and r , respectively.

The Zipf-Mandelbrot law is used for best fit search, according to the following rule:

$$s = f(r) = \frac{\alpha}{(\beta + r)^\gamma}, \quad (4)$$

where α, β, γ are parameters to be calibrated for fitting the sample under investigation.

As we will see, there is a very good compliance of the considered data with the Zipf-Mandelbrot law (see rows (a) of Table 2, and Figure 3 in Section 5). Such a property can be used to define the measure of the core of the hapaxes.

In fact, the core of the hapaxes is defined through the H index, in a similar way in which it has been introduced by [25] to evaluate scientific research. Specifically, such an index is \bar{H} when \bar{H} is the maximum number of words whose frequency is at least \bar{H} . The resulting set of \bar{H} words is the core of the hapaxes.

In practical terms, the H -index provides a formal identification of the contrast between the most frequent hapaxes and the less frequent ones. Such an indicator provides also a relevant information on the overall distribution of the hapaxes – thus, also on the way in which US Presidents have historically decided to pronounce some specific words only once in a speech – since it is strongly dependent on how ranked data are positioned in terms of their sizes. In this respect, it is worth

pointing out that a low (high) value of H stands for a large (small) distance between hapaxes which are consecutive at high ranks. This outcomes suggests that the value of H allows to explain if US Presidents focus obsessively on a few specific words to be pronounced only once (case of low value of \bar{H}) or, conversely, such an imitative behaviour does not take place.

By employing \bar{H} and the best-fit curve defined in (4), with parameters in Table 2, block (a) – justifications for choosing such estimated values will be detailed later – we are able to provide an absolute and relative measure of the core of the hapaxes. We denote such measures as \mathcal{M}_A and \mathcal{M}_R , respectively. They are defined as the area of the region below the curve in (4) delimited by $r = 1$ and $r = \bar{H}$ and as the ratio between such area and the area of the overall region, from $r = 1$ to $r = 31074$, respectively. Specifically, the absolute measure of the core of the hapaxes is

$$\mathcal{M}_A = \int_1^{\bar{H}} \frac{\hat{\alpha}}{(\hat{\beta} + r)^{\hat{\gamma}}} dr, \quad (5)$$

while the relative measure is

$$\mathcal{M}_R = \frac{\mathcal{M}_A}{\int_1^{31074} \frac{\hat{\alpha}}{(\hat{\beta} + r)^{\hat{\gamma}}} dr}. \quad (6)$$

5 Results and discussion

The results of the best-fit exercise are reported in Table 2, section (a), where one can find the calibrated parameters with the confidence intervals at 95%. The value of R^2 is 0.9971, which suggests a quite perfect compliance of the considered ranked dataset with the rank-size Zipf-Mandelbrot law. Figure 3 further supports such a result by proposing a visual inspection of the fit. A mere power (Zipf) law gives a much worse fit.

In general, one can notice some kind of hidden pattern in the data when the fit is made by a power law. This is rather usual; at low rank, it is attributed to king and vice-roys effects; at large rank, one attributes deviations to the fact that the frequency is low (=1 in our case) and because there is a plethora of quantities (different words, in our case). This is usually much complicating numerical work, for any type of theoretical and empirical fits when there is such a high density of points but with low “weight”. For information, there are more than 10000 (exactly 10088) “absolute hapaxes”; for completeness let us mention that there are 7574 “relative hapaxes” appearing twice. Recall that there is a little bit more than 31000 data points.

That is why we have run the fit reported in Figure 4, with parameters reported in Table 2, rows (c). Such a figure provides also a very intuitive preliminary interpretation of the core of the hapaxes. Indeed, the removal of the core of the hapaxes from the overall ranked list lets the fit be quite perfect; this points to the claimed king and vice-roys effect mentioned above.

| | | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\gamma}$ |
|-----|-------------------|--|----------------|----------------|
| (a) | Estimations | 6.029×10^8 | 2540 | 1.896 |
| | Conf. Interv. 95% | $(5.676 \times 10^8, 6.381 \times 10^8)$ | (2525, 2554) | (1.890, 1.902) |
| (b) | Estimations | 287.7 | 5.903 | 0.084 |
| | Conf. Interv. 95% | (281.8, 293.6) | (4.288, 7.519) | (0.080, 0.088) |
| (c) | Estimations | 4.359×10^8 | 2668 | 1.861 |
| | Conf. Interv. 95% | $(4.083 \times 10^8, 4.634 \times 10^8)$ | (2652, 2685) | (1.854, 1.867) |

Table 2: (a) Best-fit parameters of the Zipf-Mandelbrot law in Eq. (4) when all hapaxes are considered. (b) Best-fit parameters of the Zipf-Mandelbrot law in Eq. (4) for the case of the hapaxes belonging to the core. (c) Best-fit parameters of the Zipf-Mandelbrot law in Eq. (4) for the case of all the hapaxes without those belonging to the core. The ranges of the confidence intervals at 95% for the three parameters are reported in parentheses for all the cases.

We are now ready to treat the subsample given by the core of the hapaxes. By looking at the data we have $\bar{H} = 182$, i.e. there exist 182 words whose frequency is at least 182 and, simultaneously, there are not 183 words with frequency at least 183. The most frequent hapaxes are reported in Table 3 for the reader convenience. To save space, only one third of the core is shown, i.e. the most frequent 61 hapaxes.

Thus, by applying formulas (5) and (6), and by using the values listed in Table 2, section (a), a straightforward computation gives that

$$\mathcal{M}_A = \frac{\hat{\alpha}}{-\hat{\gamma} + 1} \left[(\hat{\beta} + \bar{H})^{-\hat{\gamma}+1} - (\hat{\beta} + 1)^{-\hat{\gamma}+1} \right] = 35783.9769 \quad (7)$$

and

$$\mathcal{M}_R = \frac{(\hat{\beta} + \bar{H})^{-\hat{\gamma}+1} - (\hat{\beta} + 1)^{-\hat{\gamma}+1}}{(\hat{\beta} + 31074)^{-\hat{\gamma}+1} - (\hat{\beta} + 1)^{-\hat{\gamma}+1}} = 0.0663 \quad (8)$$

Notice that the hapaxes contained in the core represents a small percentage – about 0.58% – of the entire set of words said once. However, in terms of frequencies, we have that the core is 6.63% of the overall set, as the relative measure assures. This means that a very small set of words have been selected to be said only once in a large number of speeches, with about eleven times the frequencies over the hapaxes. One can conjecture that these are rare words but purposefully intended.

To have a view of the set of the core, we report in Table 1 column (b), the main statistical indicators for the frequencies of the set of such 182 hapaxes. By exploring the core of the hapaxes itself, one can see that the frequencies of the tokens therein contained represent a sample which is well fitted by a Zipf-Mandelbrot law. Refer to Figure 5 and Table 2, rows (b) for the details. The

| Word(s) | Frequency |
|---|------------------|
| sense | 250 |
| given | 247 |
| bring, house | 240 |
| give | 239 |
| hand, themselves | 229 |
| within | 228 |
| others, therefore | 225 |
| set | 224 |
| take | 222 |
| second | 221 |
| find, full, making, since | 220 |
| among | 217 |
| again, does, | 215 |
| itself, remain | 214 |
| being, brought, done, soon, whose | 213 |
| part, protect | 212 |
| known, small | 211 |
| able, beyond, carry, friends | 210 |
| call, day, far, fellow, means, opportunity, then, Washington, while | 209 |
| course, order, single | 208 |
| essential, important, meet, reason | 207 |
| another, left, like, respect, seen | 206 |
| certain, few, necessary, possible, purpose | 205 |

Table 3: The most frequent 61 hapaxes, along with their frequencies. Other hapaxes have frequency larger than 200. Specifically, the positions/ranks between 62 and 80 have frequencies ranging in [204 - 200].

statistical goodness of fit is rather satisfactory also in this case, with $R^2 = 0.978$. Also the visual inspection suggests good compliance of the data with a Zipf-Mandelbrot law, even if some evident deviations appear (see Figure 5). Such deviations are confirmed also by the wider confidence intervals resulting in the case of the core with respect to those coming from the overall sample, see Table 2 rows (a) and (b). Hence, similarly to what happens for the entire sample case (see Figure 3), also in this case we have some points at the lowest ranks that are above the curve. Thus, the presence of the king and vice-roys effect is confirmed also when one takes such a subsample. The meaning of this outcome can be found in the presence of a few special words which have been pronounced only once in a large set of speeches. We can say that such words are *necessary, but not to be stressed*. The five lowest ranked words in the core are *sense, given, bring, house, give*. Their use is often ineluctable, but their abuse is not politically rewarding. Therefore, they appear often as hapaxes in a speech. For more details on the necessary, but not be stressed words, see the next subsection.

Figure 5 highlights also a queen and harem effect² at the highest ranks. Indeed, the hapaxes of the core which appear more rarely in the US Presidents' speeches are below the best fit curve. This outcome confirms the linguistic evidence that the less frequent hapaxes of the core are relatively less scattered than the more frequent ones. Furthermore, the elements of the core with highest ranks form a subset of *identically pronounceable words resulting to be hapaxes*; specifically, such words form a collection of tokens which have been used in the same number of occasions by the US Presidents. The queen and harem effect appears because the set of such identically pronounceable words used as hapaxes has a particularly large cardinality. In noting this, we provide further confirmations that the US Presidents use words as devices for communication strategies, having a clear idea on the proper terminologies – always the same – to be used only once in specific situations. This goes in the direction of stating a substantial tendency of the politicians to imitate the past when delivering a talk, as already discussed in the paper.

The presence of king and vice-roys plus queen and harem effects distort the best fit curve, which then is constantly below a group of middle-rank hapaxes of the core (see again Figure 5). Thus, such a special set of elements of the core – denoted hereafter as *regime of the highly frequent middle-rank hapaxes* – has a behavior which is analogous to that of the identically pronounceable words used as hapaxes. In fact, the existence of the regime of the highly frequent middle-rank hapaxes is due to the properties of such words to be pronounced as hapaxes about the same number of times. Specifically, hapaxes in this set have the same size or there are no noticeable changes in size when the one moves from a rank to the subsequent one. Therefore, one can offer an interpretation of the regime of the highly frequent middle-rank hapaxes which is similar to that of the set of identically

²Which is the respective of king and vice-roys but manifested at the highest ranks

pronounceable words employed as hapaxes.

Notice that the cardinality of the sample set is able to affect the goodness of fit; in particular, in some circumstances, one can claim that larger cardinality leads to less scattered data.

The hapaxes in the core produce a king (the word *sense*, with frequency 250) and 181 vice-roys effect. Indeed, once the core is removed from the sample, one obtains a perfect fit through a calibrated Zipf-Mandelbrot law, because of the removal of the deviations at the low ranks (compare Figures 3 and 4). Similar deviations at low rank can be found in studies on other types of data, e.g. city size, or co-author distributions (see e.g. [37, 5, 12]). In the case of core removal, the goodness of fit remains quite perfect, with $R^2 = 0.9965$. The best fit parameters can be found in Table 2, rows (c), along with the related confidence intervals.

5.1 Implications of the analysis: a discussion of some salient hapaxes

As a preliminary premise, we argue that the most evident implication of our proposal lies in the identification of relevant hapaxes (those in the core) whose meaningfulness has to be discussed also in the light of their role in the context.

Indeed, it is important to claim that a full and complete description of the connection between the single words in the core and the US Presidents communication strategy can be properly discussed only by involving a semantic and local analysis of the tokens – along with their contextualization. In saying so, we claim that this paper can be viewed as the presentation of a promising methodology and the starting point of future studies in the linguistic arena.

We here provide some suggestions leading to a more detailed interpretation of the results of the analysis. In particular, a discussion of some of the most frequent hapaxes resulting from the analysis is reported here to show the potentiality of the presented approach. Such a discussion proceeds in two steps: first, we give some general insights on the content of the core of the hapaxes, with a special attention to some classes of such tokens; second, we provide a detailed analysis of some specific highly recurrent hapaxes, along with a discussion on their contextualization in noticeable speeches. We have also implemented a specific poly-grams³ analysis to explore the context in which the hapaxes occur. The phrases reporting the terms of interest are reported and commented in the next rows.

Still in the line of highlighting the implications and the applications of our approach, we also report a discussion of some hapaxes appearing strictly once in the entire corpus.

In the core, one can find the following terms: *opportunity*, *purpose*, *confidence*, *progress*. Such words suggest a positive reading of the subject of the speeches. Since they are hapaxes, one can

³A poly-gram is a contiguous sequence of n words from a speech transcript.

reasonably argue that they are associated to a substantially negative situation – hence, they are pronounced only once – which the Presidents try to revert into a positive perspective. This communication strategy has appeared several times in the history of the US Presidents speeches. For example, “*And our neglected inner cities will see a rebirth of hope, safety, and opportunity.*” is the phrase pronounced by Donald Trump in the Address to Joint Session of Congress on February 2017.

Another class of nouns collects duty, efforts, honor, justice, responsibility, effort, others, friends, respect, order, safety, faith, hope, hands. Here the Presidents appeal to the sentiments of the auditors and call for empathy and political support. Also in this case, intuition suggests that the content of the messages containing such hapaxes was related to troubling situations. In this worrying context, the brief reference to the worthiness of the citizens can be viewed as an effective communication device against pessimism. For example, from the aforementioned talk stated by Trump on February 2017: “Also with us are Susan Oliver and Jessica Davis. Their husbands, Deputy Sheriff Danny Oliver and Detective Michael Davis, were slain in the line of duty in California. They were pillars of their community. These brave men were viciously gunned down by an illegal immigrant with a criminal record and two prior deportations. Should have never been in our country.” or the State of the Union Address from Bill Clinton on February 1997, where the President said: “We have much to be thankful for. With four years of growth, we have won back the basic strength of our economy. With crime and welfare rolls declining, we are winning back our optimism, the enduring faith that we can master any difficulty. With the Cold War receding and global commerce at record levels, we are helping to win an unrivaled peace and prosperity all across the world.”

In a similar context, the introduction in the speeches of protect, respect, support as hapaxes points to the efforts spent by the Presidents in defending the auditors, even in presence of adversities. For example, President Obama in the Speech on the Strategy in Afghanistan and Pakistan in December 2009 said: “For the first time in its history, the North Atlantic Treaty Organization invoked Article 5 – the commitment that says an attack on one member nation is an attack on all. And the United Nations Security Council endorsed the use of all necessary steps to respond to the 9/11 attacks. America, our allies and the world were acting as one to destroy al Qaedas terrorist network and to protect our common security.”

We now point our attention to a set of adjectives which belong to the core of the hapaxes: essential, important, necessary. Such tokens can be associated to the political will of the speakers when referring to the ineluctability of an intervention which presents potentially negative outcomes. For example, President Obama talking to the UK Parliament, just after mentioning the joint intervention in Libya in 2011, said: “Our action - our leadership - is essential to the cause of human dignity. And so we must act - and lead - with confidence in our ideals, and an abiding faith in the

character of our people, who sent us here today.”

The classes of hapaxes of the core introduced above allow to have a general view of the way in which US Presidents and their collaborators have created the official speeches. Some communication strategies appear with a certain degree of recurrence, and this goes in the direction of explaining the presence of an imitative behavior among the Presidents.

We are now ready to clarify some more specific regularities in the Presidential talks by analyzing the patterns of some peculiar hapaxes.

Let us start checking which Presidents have used the most common hapax *sense* (see Table 4, left box) with a specific focus to the reference context. Four Presidents (John Adams, Martin Van Buren, Zachary Taylor and Rutherford B. Hayes) noticeably used “sense” as hapax. They employed it in circumstances of celebrations or very formal contest like Annual Messages, Proclamations or Veto announcements. The refine rhetoric used at the time of the aforementioned Presidents (1797 - 1883) jointly with particularly important aforesaid appointments of Presidents’ political agenda, create the ground for the use of the hapax “sense” to qualify feelings, with references to a common perception of something. For example, Chester A. Arthur, in “Veto of River and Harbors Act ” of August, 1882 said: *‘It is not necessary that I say that when my signature would make the bill appropriating for these and other valuable national objects a law it is with great reluctance and only under a sense of duty that I withhold it.’*. Chester A. Arthur has used “sense of duty” in the circumstances of a Veto Message, when he was trying to convince his audience about his arguments⁴, but the Congress has overridden the veto and the legislation was approved.

John Adams, Martin Van Buren, Zachary Taylor and Rutherford B. Hayes have evoked sentiments like *‘sense of national honor, dignity, and independence ’* (John Adams; December 8, 1798: Second Annual Message), *‘the good sense and patriotism’* (Martin Van Buren; December 3, 1838: Second Annual Message to Congress) or *‘sense of the duty’* (Rutherford B. Hayes; December 1, 1879: Third Annual Message) referring to specific common feelings. From 1900 on, the bigram “common sense” has been employed 29 times out of 250 times that “sense” appears as hapax. The Presidents who have evoked most frequently “common sense” are Ronald Reagan and Franklin D. Roosevelt. The former points to the common sense in formal contexts like State of the Union Addresses of January, 1988 and February, 1986, while the latter appeals to common sense during his famous fireside chats (Fireside Chat 5: On Addressing the Critics, Fireside Chat 10: On New Legislation, Fireside Chat 15: On National Defense). Bear in mind that the fireside chats have been instituted by the President to have a colloquial level of communication with citizens. Here we report some words pronounced in the conclusive part of the “Fireside Chat 10: On New Legislation” of October, 1937: *‘The common sense, the intelligence of the people of America agree with my statement that*

⁴www.u-s-history.com/pages/h735.html

“America hates war. America hopes for peace. Therefore, America actively engages in the search for peace.” ¹. This part of the discourse was about peace. However, the President mentions “common sense” once, to deliver the (hidden) message that US citizens should exhibit responsibility and unity in the terrible case of non avoidable war (that was the case, indeed, as the President probably guessed. Remember that Hitler had re-militarized the Rhineland in March 1936, just before the speech). Thus, hapax could be a part of a bigger rhetoric framework adopted by Franklin D. Roosevelt to prepare the public opinion for future military actions without explicitly mentioning them. Substantially, Presidents are aware that the bigram “common sense” might be efficiently used to introduce wisdom and sustainability criteria in military and engineering contexts.

Concluding, the presence of “sense” as the *king* of the hapax list can be justified by assuming that the speakers\writers have paid attention in referring to any type of common sense when dealing with public communications, even if without stressing it too much. Indeed, when a public speaker is invoking a shared sense of something, he has to bear in mind that, what is considered common sense for him or for a certain sub-community of auditors, is not common sense for others, so using certain hapax locutions just once, he can send a specific message to a particular group (characterized by the appropriate sensibility needed to receipt that particular message) or he is trying to keep people feeling the sense that he is appealing to, in order to inspire a specific collective behaviour, e.g. see [29].

A second interesting hapax is *bring*. From Table 4 right block, emerges that Presidents Richard Nixon, Harry S. Truman, Chester A. Arthur and John Tyler are the Presidents most often utilizing the hapax “bring”. The aforementioned Presidents did not employ “bring” in colloquial situation, at least in the speeches stored in this dataset. The hapax occurs in formal and relevant political events to manifest the immediate need of an action, or to describe the effects of a full willingness to act in a certain direction. In particular, the hapax “bring” has a relevant use in relation to tensions, conflict or war. For example, in “Address to the Nation on Presidential Tape Recordings” of April 1974, Richard Nixon said *‘These conversations are unusual in their subject matter, but the same kind of uninhibited discussion and it is that the same brutal candor is necessary in discussing how to bring warring factions to the peace table or how to move necessary legislation through the Congress.’* referring to the political struggle started with the Watergate case. Another salient example is the speech of September 1945, “Announcing the Surrender of Japan” stated by Harry S. Truman, where he said *‘No victory can bring back the faces they longed to see.’* referring to the deaths of that war; or the speech of September 1948, “Whistlestop Tour in Trenton, Missouri” during which Harry S. Truman said: *‘There is one thing I want to bring home to you.’* This comment can be extended to all the other Presidents that have used “bring” as hapax: indeed the word “war” occurs 37 times within the phrases in which “bring” appears.

The behaviour of “bring” as hapax along the Presidents is different from the previous one, (compare columns in Table 4). The “bring” occurrence is more homogeneously distributed. Consequently, it is more difficult to grasp dramatic changes along the presidencies, but we can note that the hapax “bring” has a longer life, being used from November 6, 1792, by George Washington in his “Fourth Annual Message to Congress” to very recent speeches.

The hapax legomenon *house* has a peculiar regularity in occurring, because it is commonly used in the introductory statements of many ceremonial moments of the US political agenda. For example, the introductory forms ‘To The Senate and House of Representatives’ and ‘Fellow Citizens of the Senate and of the House of Representatives’ are the source of the usage of “house” as hapax for 53 and 26 times respectively (a poly-grams comparison has been run to figure it out). These forms were commonly used between January, 1790 (“First Annual Message to Congress” of George Washington) and December, 1932 (“Fourth State of the Union Address” of Herbert Hoover). The utilization of such locutions has originated from the fact that during the aforementioned period, most of the messages were not personally stated by the Presidents, but they were spread in written form (Thomas Jefferson, Woodrow Wilson and Franklin D. Roosevelt are the unique that have not respected this common practice). According to [36], Franklin D. Roosevelt established the personal appearance as a permanent tradition with his 1934 State of the Union Message. After that talk, the so called “Presidents Annual Message to Congress” starts to be known as “State of the Union Address”. Those points are confirmed by our findings. The two introductory locutions listed above were employed mainly for Annual Messages and State of the Union Addresses until 1932, but later, under the Franklin D. Roosevelt Presidency, they stopped to occur. After that, the use of “house” is mostly associated to the bigram “white house” which justifies the use of the hapax for 64 times out to 240. In the light of this evidence, we point out that the employment of “white house” as hapax has been intensified after Harry S. Truman, whose Presidency is the one during which the White House has been restored⁵. Furthermore, thank to a visual inspection of the context in which “white house” has been used, it results that the Presidents refer to it for indicating more than a mere physical place (it is a case of *metonymy* utilization). They address to the White House for pointing to the residence of the US most representative political institution. For example, on March 21, 2013, during the “Address to the People of Israel”, Barack Obama said: *‘Just a few days from now, Jews here in Israel and around the world will sit with family and friends at the Seder table, and celebrate with songs, wine and symbolic foods. After enjoying Seders with family and friends in Chicago and on the campaign trail, Im proud that I’ve now brought this tradition into the White House. I did so because I wanted my daughters to experience the Haggadah, and the story at the center of Passover that makes this time of year so powerful’*. Such a speech highlights

⁵<http://www.whitehousemuseum.org/special/renovation-1948.htm>

the importance of having certain events into the White House, and the hapax implicitly states the relevance for all the Americans.

Additionally, to do the best for providing insights on the worthiness of the analysis of the words pronounced only one time, we now discuss some hapaxes appeared just once in the whole corpus. Their meaning is relevant for the political context even if they are not belonging to the core of hapaxes.

The names *Dostoyevsky*, *Kandinsky*, *Scriabin*, *Uzbek*, *Alisher Navoi*, *Boris Pasternak* and *Zhivago* have been stated by the President Reagan on the May 31, 1985 during his ‘Address at Moscow State University’. They are good examples of meaningful hapaxes. Indeed, they occurred all together when the President Reagan was trying to emphasize his appreciation and familiarity with the Russians’ culture. He has used them for giving strength to a speech mostly dedicated to “freedom” and “truth” just few month before the fall of the Berlin wall. However, it was a local phenomenon confined in a specific situation. Differently, the cases of “sense”, “bring” and “house” represent something of a more general contextualization; in fact, they belong to the core of hapaxes. This is expected because the threshold calibrated on the H-index helps in identifying a list of hapaxes whose messages are constantly present across the talks. Summarizing, our proposal catches unit of information (words) due to their degree of systematic singular usages. Consequently, it is easier to highlight the delivered latent messages; this leads to the reading of the US political communication history under a different perspective.

5.2 Extension to poly-grams: some remarks

A relevant theme to be discussed is the extension of the analysis to poly-grams said only once in each speech. Indeed, bi-, tri- and, in general, poly-grams might turn out to be useful for grasping further information on the political speeches. However, there is no room here to face this aspect in an exhaustive way. We here elaborate on this relevant problem.

First of all, we have arguments for thinking that the exploration of poly-grams is quite much demanding from an analytical point of view. In fact, the one word case can be treated by considering a token as a unique set of consecutive letters, and two tokens as divided by blank spaces or punctuation. Bi-grams are two consecutive tokens considered together, and the distinction rule applied for one word is not longer applicable in this case. Therefore, the shortest meaningfulness unit of analysis has to be the one-word token, hence leading to the evidence that the main driver of the sense of a poly-gram would remain the same we are considering here for one word. Moreover, the number of bi-grams said only once should evidently be much higher than the number of hapaxes, and the number of tri-grams said once should be greater than the case with two words.

Such a tendency goes on till the point in which the number of words of the poly-gram is small enough. Indeed, in the extreme case of one hundred words poly-gram in a speech with one hundred words, then the entire speech can be seen as the only poly-gram said only once. However, when aggregating over all the speeches, the growing number of cases to be treated for bi- or tri-grams may lead to an extremely complex computational procedure, specially for large corpus, and this can turn out to be a severe drawback of the methodological analysis of the problem.

By a completely different perspective, a relevant issue to be carefully considered concerns the meaning of the poly-grams. One-grams are words with certain meaning; they can be considered, as argued here, as if bringing purposeful ideas. The meaningfulness of the poly-grams is somewhat questionable. One can have sequences of words which are forced by grammatical and syntactical constraints, or poly-grams whose logical sense cannot be appreciated when taken out from the context. As an example, if we consider the bi-gram "America is", the selected word is "America", while the term "is" is a trivial grammatical constraint. The President might also select "Our country is". The real difference lies in the choice between "America" and "country". In the former case, attention is paid explicitly to America, while in the latter one the President points to a more general term like "country". It is also worth to observe that if "America" is a hapax, then "America is" is a bi-gram said only once. The converse is not true in general, and even if "America is" is a bi-gram said only once, the word "America" is not necessarily a hapax. So, even if different information are captured, there will be the need of a reading phase to detect those poly-grams really referring to a recurring marginal topic in the US President speeches.

As a further consideration, one can easily see that there is a critical set of words which are associated to poly-grams said only once in a specific speech. The number of words in such a critical set does not increase as the length of the speech decreases. As an intuitive example, if we take a sentence with ten words, most likely all the bi-grams are never repeated more than one time. In this respect, we observe that the length – in term of words – of a generic US Presidents' speech is rather small. Thus, one may likely have that all the poly-grams composed by a small number of consecutive words are said only once. Such remarks suggest that the outcomes of the study of the poly-grams can present some sources of biases. We will go back to this point in the next Section, when discussing potentially interesting research topics.

6 Conclusive remarks

This paper faces the challenging theme of exploring part of the content of the official political speeches with an innovative method. The paradigmatic case of US Presidents' talks serves as a guide. We start from the premise that official speeches are carefully written because the messages

carried out are highly influential. Each talk contains information to be delivered to some audience and aimed at the entire society. Thus, words are tactically selected with care, depending on the situation.

We are interested in the hapaxes of each speech, which are relevant units of information in the Presidents communication strategy. In fact, one can observe some recurrent hapaxes in the corpus. They have consciously pronounced / selected only once in several occasions and by several Presidents. The relative rarity of these is thought to be intentional, sometimes appearing as new (or astute) words, implying the President modernity, elitism, and wide knowledge. Anyway, we consider the approach here designed as promising for studying even much larger corpora that spreads across many years.

If appropriately merged and ranked, hapaxes show regular paths and can be successfully fitted by the Zipf-Mandelbrot law. Moreover, there is a privileged set, the core hapaxes, here defined through the introduction of a Hirsch-based threshold.

We have shown that a small number of words have been pronounced once per speech several times in official communications. This confirms and lets us understand the presence of common messages and arguments in the historical paramount view of the US Presidents' interventions. This list represents the core of the hapaxes. Such a core can be interpreted as those words which strike a point, even though they are rarely used within each text of the corpus.

We have also shown that the core has a structure similar to the one of the overall sample, with a compliance with a rank-size law of Zipf-Mandelbrot type. Moreover, the core is also responsible for deviations of the overall set of hapaxes from the best Zipf-Mandelbrot curve. In this, king and vice-roys effects are detected.

The analysis of some relevant hapaxes is presented in Section 5.1, to illustrate the implications of the new approach. There we have looked for the contexts in which certain hapaxes occurred, justifying their presence. This is fundamental to understand the potential of our approach.

From Section 5 the ability of our method in responding to the research questions clearly emerges. Indeed, the hapax legomena occurring in a collection of texts can be huge and their meanings are remarkable, as already stressed above. One should go through each of them manually and understand the context in order to decide which one are connected to a global phenomenon and which are linked to a local one.

It is important to notice that this study cannot offer a systematic analysis of all the hapaxes which occurred a few times (and, in particular, just once) in the whole corpus of the US Presidents' speeches, because their contextual meaningfulness in the respective speeches may lead to a very historic / social specific context, a spurious results or to a somewhat questionable informative content. In this respect, it is worth noting that in defining the core of hapaxes and a procedure

to determine it, we filter out tokens until we get a list of words which have been regularly used by the Presidents; therefore, the words in the core of the corpus have passed a selection process through years of usages and political phases, sometimes ending up to be pronounced just once per speech. On the other hand, certain hapaxes occurring just one time in the corpus may be part of a very local event. To be able to capture it, a reading inspection performed by an expert like a philologist rather than an automated procedure is required; this is well-beyond the target of the present paper. Differently, in the case of those words belonging to the core, the likelihood to be relevant is much higher.

Finally, the presented analysis is fully reproducible in different contexts and is particularly useful when a researcher is facing a large collection of documents coming from a uniform environment and difference sources. Indeed, if s/he wants to easily access the presence of marginal topics or remarkable semantic outliers, the analysis of the core of the hapaxes can be of special usefulness (see [23] as example of an extensive analysis of hapaxes and the justification of such an investigation). Furthermore, our procedure supports the assessment of a cross-document view, which is helpful for the identification of latent marginal keywords that may point to a *fil rouge* between texts.

6.1 Future research

We observe that the present study represents a further step towards the methods to investigate corpora; it is a step ahead in the comprehension of the political speeches. Furthermore, such a tool can be considered as helpful for making decisions on the words choice to deliver certain information, starting from messages that have been stated by US Presidents during the US history. In this respect, some final remarks on the obtained results point to interesting future research.

For what concerns the kinds of speech which would deviate from the single exponent rank-size law patterns, one should consider that the speeches present “multifractal aspects”. In this respect, see e.g. [60, 3, 4, 19].

The analysis of how such speeches should be received has been already a little bit discussed in [2, 4, 76]. In this context, it would be interesting to use readability tests, readability formulas, or readability metrics for evaluating the readability of President speeches; this should be (usually made) by observing punctuation, and counting syllables, words, and sentences. One might extend such criteria looking for “word correlations”, and in the present cases for the position of the hapax in the speech. Some readability formulas refer to a list of words graded for difficulty. For what we have seen, the hapaxes are not “difficult to understand” words (see the Section 5.1 and Table 3). Of course, one could also debate about the difference between “readability” and “understanding”; moreover reading and hearing concern two different senses. Here we assume them to be rather

identical.

In addition, we are able to assess a cross-document view through low-frequency hapaxes, which may point to connection among speeches. Such a challenging research theme might stimulate work for future research, mainly in the linguistic arena of information science.

In the same line, the study can be extended to the case of poly-grams said only once by taking into full consideration the presence of some correlations among them, to remove crucial points as those raised in Subsection 5.2.

From the theoretical perspective, a promising research direction consists in formalizing the generative process for the cross-texts hapaxes' rank-size behaviour. In the environment of the rank-size procedures, the assessment of the model behind the final distribution of the ranked data is *per se* of scientific relevance. We are not aware of scholars dealing with the connection between the underlying stochastic process and the ranked phenomenon. The approach for the construction of a probabilistic model related to a rank-size law is grounded on the interpretation of the resulting rank-size distribution the outcome of a stochastic process. In a preferential attachment context, the idea is to define a step-wise procedure in the framework of the urn and in presence of rules stating the addition of balls in the urn at every step, as in Polya's process. An example is found in [7], where a rank-size law was also discovered. The Polya urn stochastic structure procedure can be meaningful and valid here – under the obvious requirement that the asymptotic distribution of the stochastic process represents a statistically significant approximation of the rank-size law.

Under the point of view of the linguistic field, we acknowledge the evident discrepancies between political and other types of texts – like scientific ones. In so doing, we aim at providing some intuitive insights on the expected outcomes of future rigorous studies.

Indeed, scientific speeches can be of two very different categories. By one side, scientists deliver speeches to illustrate the content of their studies to the scientific community. Appropriate occasions for this type of talks are conferences and seminars at Academic Institutions; by the other side, a scientist can pursue the scope of divulging her/his scientific achievements to people not involved in the field. In this case, the talk can be delivered during a conference press or in non scientific meetings open to the public. In this latter category, one can find also the situation in which scientists inform politicians and policymakers on the results of research activities having an impact on the society – hence, calling for regulation and political actions.

The scientific speeches addressed to a scientific community have a structure which is structurally biased by their contents. Indeed, researchers are constrained to employ some specific words when referring to peculiar concepts, to maintain the level of the text as rigorous as possible. Basically, synonyms do not appear often in pure scientific talks. As an example, we can mention the derivative of a (regular enough) function. It is possible to refer to the derivative by means of a formula,

but scientists do not have a different term to denote such a mathematical concept. Therefore, when the content of the study includes the derivative of a function, the speaker is forced to employ always the word *derivative* when referring to it.

A completely different perspective is associated to scientific talks aiming at disseminating a research achievement. In this case, the authors of a scientific study are obliged to avoid technical terms; hence, they can take the advantage of expressing scientific concepts by means of words of common use. In so doing, researchers are able to properly build their speeches in the light of a pre-selected communication strategy.

The arguments expressed above suggest that political speeches are quite similar to scientific speeches, when the latter ones are of divulging type. If the scientific speech has a purely academic and technical nature, then the selection of a peculiar linguistic structure or the decision of employing some particular hapaxes is quite out of the range of the possibilities of the speaker.

In this respect, it is important to mention [59], where the Author elaborates on the communications between politicians and scientists. Specifically, the former category is claimed to serve as a target audience for divulging the outcomes of scientific research. Indeed, as pointed out above, it is crucial that politicians understand the most relevant scientific contexts. The quoted paper clearly illustrates that scientific divulgation and political messages should follow similar communication patterns in order to obtain such a desired target.

Thus, one can argue that researchers have to imitate politicians in terms of communications, if they aim at letting their studies be understandable by the policymakers. Of course, this represents a difficult task for scientists.

To conclude, a detailed study is expected to offer a high level of discrepancy when comparing the linguistic structures of scientific speeches and political ones. Such deviation would tend to reduce when researchers pursue the target of stimulating the action of the politicians. In terms of hapaxes, the scientific speeches aimed at divulging science should present a core of hapaxes made of non-veiled topics but rather the core will contain technical words demanded by the context. Therefore, to produce a meaningful analysis with scientific talks, one should get a homogeneous corpus, for example the Nobel Prize speeches.

Under a more general perspective – but still in the linguistic arena – one can advance some conjectures representing additional noticeable topics for future research. Such conjectures provide also a motivation for the present study.

First, we conjecture that a word said only once is not always delivered to all, but it might be pronounced to reach a qualified group of auditors who are able to grasp the message. In this respect, evidence suggests that a word pronounced several times rings the bells of the widest part of the audience, and sometimes the comprehension of the mass is viewed as a negative outcome of the

speech; second, we conjecture that hapaxes might reflect concepts whose explanation is absolutely necessary, but associated to situations that the President recalls reluctantly to the audience. In this sense, in order not to be attracting too much the attention of the auditors to a specific argument – whose mention is, however, necessary – the Presidents might wish to pronounce some words only once; third, we conjecture that hapaxes might be related to situations and concepts that should not be included in a speech and, despite this fact, are intentionally mentioned by the President. Intuitive examples of such cases are the ones related to diplomatic accidents. In such situations, the annoyance of a President cannot be strongly stressed, in the light of maintaining excellent relationships with institutions and commercial partners. However, the President cannot be silent in the regard of an insult or of an improper behaviour, in order to not lose credibility. The hapax is the right mean for these types of situations.

Furthermore, the analysis of the frequencies of the hapaxes over a large set of speeches – which is exactly our methodological ground here – provides additional insights. Indeed, all the conjectures listed above should be interpreted in the more complex environment of several Presidents and occasions. This remark leads to some (further) research questions, that have to be treated with care in future studies. In particular, it is worth noting that largely pronounced hapaxes raise a question about the presence of recurrent themes, treated by the President(s), also in the light of the conjectures listed above. Furthermore, the recurrence of a hapax raises a question about the presence of an imitative behaviour of the Presidents to their predecessors. Substantially, there is evidence of the will to learn from the past in the communication strategies. It is also important to point out that the employment of specific hapaxes by several Presidents raises a question about the existence of a sort of linguistic code. Finally, since hapaxes are rather rare words, one can conjecture that they are taken from a dictionary black box, only containing the never used words. Some comments on these aspects are contained in Subsection 5.1. However, our conjectures and research questions can be seen as a relevant ground for proceeding in the analysis of the hapaxes in the considered corpus of the US Presidents' speeches.

This said, we point out that the analysis of the hapaxes here carried out provides a wide number of research questions, and that is exactly the philosophical target of our research activity.

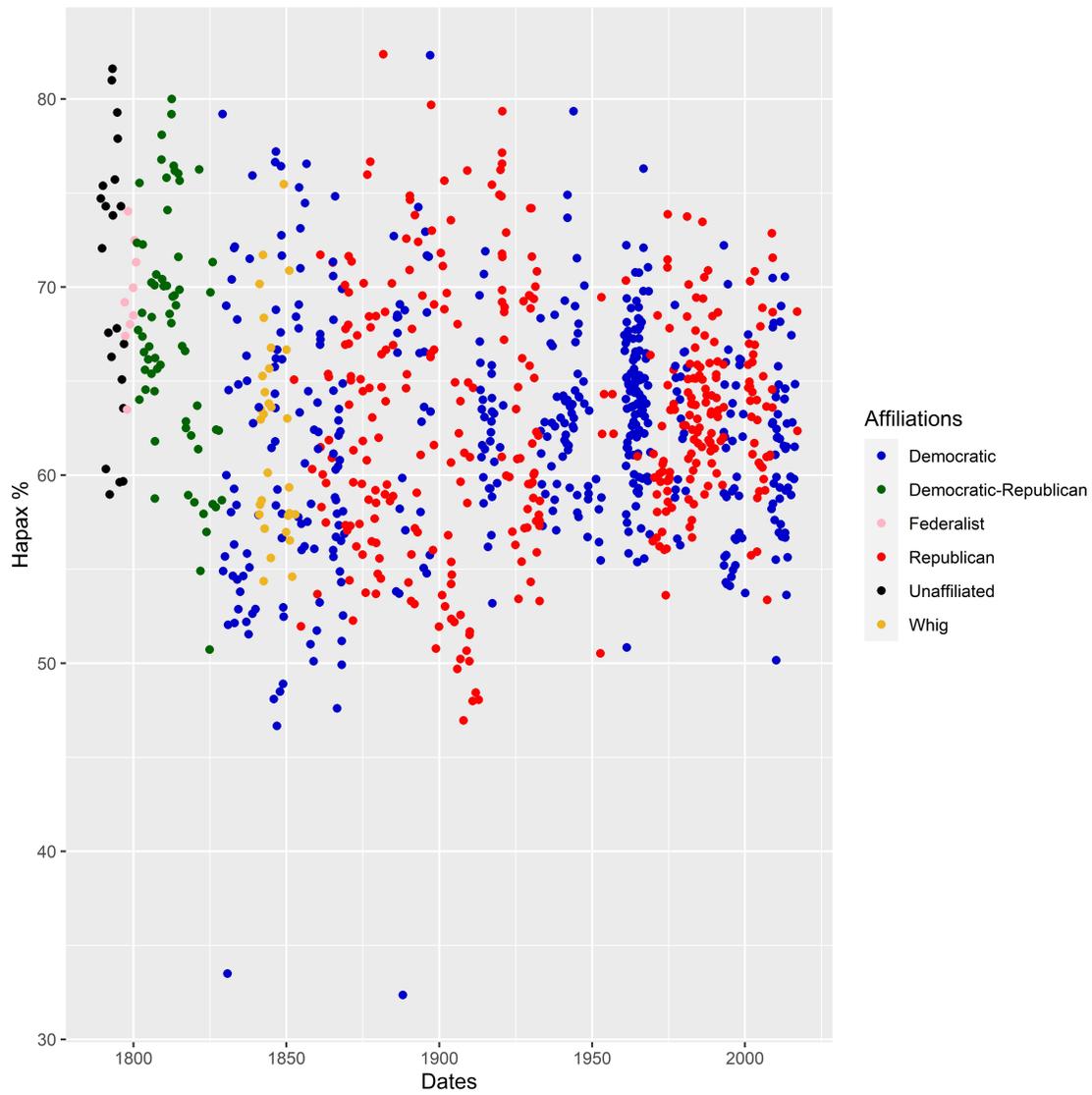


Figure 2: Percentage of hapaxes per speech ranked in chronological order. The colours represent the speakers' affiliations; they are reported for the sake of clarity.

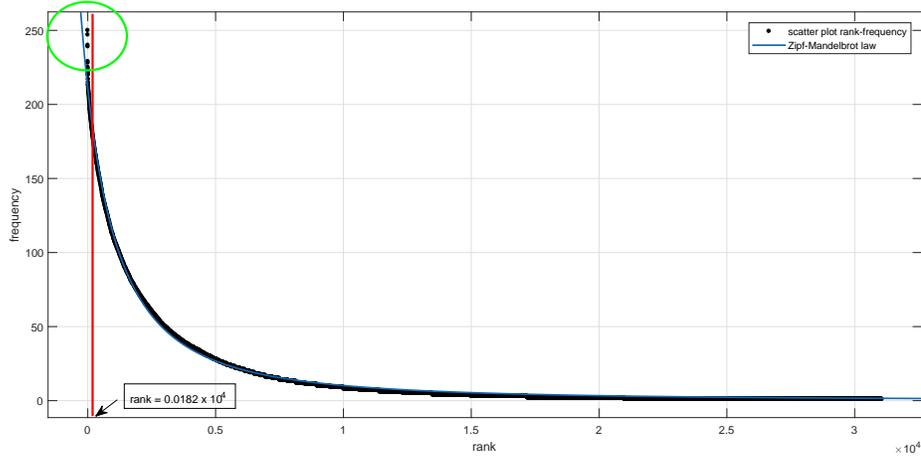


Figure 3: Best-fit curve, according to equation (4) and calibrated parameters in Table 2, rows (a). The scatter plot of the original sample is juxtaposed for a better comparison; the agreement is very good; data and fits are hardly distinguishable from each other. Notice the slight deviations at low ranks (green circle in the Figure), suggesting the presence of king and vice-roy effects (see e.g. [12]). The red vertical line points to $\bar{H} = 182$, which delimits the core of the hapaxes.

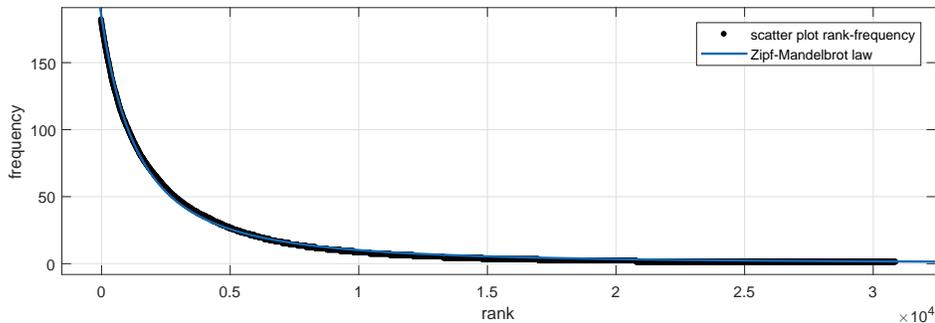


Figure 4: Best-fit curve, according to Eq. (4) and calibrated parameters in Table 2, block (c), for the case of the hapaxes excluding the core. The scatter plot and the fitted curve are not distinguishable. The deviations at the low ranks shown in Figure 3 do not appear, thus leading to the statement of the presence of king and vice-roys effects for the elements of the core in the respect of the overall sample.

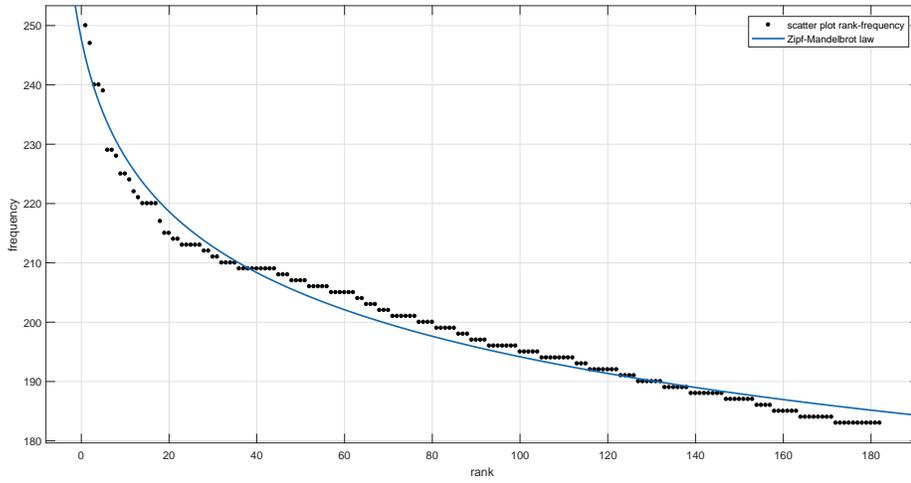


Figure 5: Best-fit curve, according to Eq. (4) and calibrated parameters from Table 2, rows (b) for the case of the hapaxes in the core. The scatter plot of the original sample of the core is also shown for comparison purposes; the agreement is visually good.

| Presidents | sense [%] | tot. speech | Presidents | bring [%] | tot. speech |
|-----------------------|-----------|-------------|-----------------------|-----------|-------------|
| John Adams | 66.67 | 9 | Richard Nixon | 39.13 | 23 |
| Martin Van Buren | 60.00 | 10 | John Tyler | 38.89 | 18 |
| Zachary Taylor | 50.00 | 4 | Harry S. Truman | 36.84 | 19 |
| Chester A. Arthur | 45.45 | 11 | Chester A. Arthur | 36.36 | 11 |
| Rutherford B. Hayes | 43.75 | 16 | Gerald Ford | 35.71 | 14 |
| William Taft | 41.67 | 12 | Andrew Johnson | 35.48 | 31 |
| Ronald Reagan | 36.84 | 57 | George H. W. Bush | 35.00 | 20 |
| Barack Obama | 36.00 | 50 | Bill Clinton | 34.21 | 38 |
| William McKinley | 35.71 | 14 | Abraham Lincoln | 33.33 | 15 |
| John F. Kennedy | 34.15 | 41 | Calvin Coolidge | 33.33 | 12 |
| Dwight D. Eisenhower | 33.33 | 6 | Dwight D. Eisenhower | 33.33 | 6 |
| George Washington | 33.33 | 21 | Jimmy Carter | 33.33 | 18 |
| Warren G. Harding | 33.33 | 18 | John Adams | 33.33 | 9 |
| Franklin D. Roosevelt | 32.65 | 49 | Ulysses S. Grant | 31.25 | 32 |
| James Monroe | 30.00 | 10 | Barack Obama | 30.00 | 50 |
| Bill Clinton | 28.95 | 38 | James Monroe | 30.00 | 10 |
| Lyndon B. Johnson | 28.79 | 66 | Martin Van Buren | 30.00 | 10 |
| Gerald Ford | 28.57 | 14 | Franklin D. Roosevelt | 28.57 | 49 |
| James Buchanan | 28.57 | 14 | Millard Fillmore | 28.57 | 7 |
| George W. Bush | 28.21 | 39 | George W. Bush | 28.21 | 39 |
| James K. Polk | 28.00 | 25 | James Madison | 27.27 | 22 |
| Woodrow Wilson | 27.27 | 33 | Ronald Reagan | 26.32 | 57 |
| Andrew Jackson | 26.92 | 26 | Lyndon B. Johnson | 25.76 | 66 |
| Herbert Hoover | 23.33 | 30 | Zachary Taylor | 25.00 | 4 |
| Richard Nixon | 21.74 | 23 | Woodrow Wilson | 24.24 | 33 |
| Benjamin Harrison | 21.05 | 19 | George Washington | 23.81 | 21 |
| Franklin Pierce | 20.00 | 15 | Thomas Jefferson | 20.83 | 24 |
| George H. W. Bush | 20.00 | 20 | Franklin Pierce | 20.00 | 15 |
| Grover Cleveland | 17.24 | 29 | John F. Kennedy | 19.51 | 41 |
| Calvin Coolidge | 16.67 | 12 | Andrew Jackson | 19.23 | 26 |
| John Tyler | 16.67 | 18 | Rutherford B. Hayes | 18.75 | 16 |
| Andrew Johnson | 16.13 | 31 | Grover Cleveland | 17.24 | 29 |
| Harry S. Truman | 15.79 | 19 | Herbert Hoover | 16.67 | 30 |
| Millard Fillmore | 14.29 | 7 | James K. Polk | 16.00 | 25 |
| James Madison | 13.64 | 22 | James Buchanan | 14.29 | 14 |
| Abraham Lincoln | 13.33 | 15 | John Quincy Adams | 12.50 | 8 |
| John Quincy Adams | 12.50 | 8 | Benjamin Harrison | 10.53 | 19 |
| Ulysses S. Grant | 12.50 | 32 | William Taft | 8.33 | 12 |
| Theodore Roosevelt | 9.09 | 22 | William McKinley | 7.14 | 14 |
| Thomas Jefferson | 4.17 | 24 | Warren G. Harding | 5.56 | 18 |

Table 4: The percentage of the speeches per President that contain the word “sense” and “bring” (respectively left and right boxes) as hapaxes. The sub-tables are ranked according to “sense [%]” and “bring [%]” respectively.

References

- [1] Ausloos, M. (2008). Equilibrium and dynamic methods when comparing an English text and its Esperanto translation. *Physica A: Statistical Mechanics and its Applications*, 387(25), 6411-6420.
- [2] Ausloos, M. (2010). Punctuation effects in English and Esperanto texts. *Physica A: Statistical Mechanics and its Applications*, 389(14), 2835-2840.
- [3] Ausloos, M. (2012). Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. *Physical Review E*, 86(3), 031108.
- [4] Ausloos, M. (2012). Measuring complexity with multifractals in texts. Translation effects. *Chaos, Solitons and Fractals*, 45(11), 1349-1357.
- [5] Ausloos, M. (2013). A scientometrics law about co-authors and their ranking: the co-author core. *Scientometrics*, 95(3), 895-909.
- [6] Ausloos, M. (2015). Coherent measures of the impact of co-authors in peer review journals and in proceedings publications. *Physica A: Statistical Mechanics and its Applications*, 438, 568-578.
- [7] Ausloos, M., & Cerqueti, R., (2016). A universal rank-size law. *PLoS ONE*, 11(11), e0166011.
- [8] Ausloos, M., Nedic, O., Fronczak, A., & Fronczak, P. (2016). Quantifying the quality of peer reviewers through Zipf's law. *Scientometrics*, 106(1), 347-368.
- [9] Bauer, L., & Laurie, B. (1983). *English word-formation*. Cambridge University Press.
- [10] Bauer, L. (2001). *Morphological productivity*, Vol. 95. Cambridge Studies in Linguistics.
- [11] Borko, H. (1968). Information science: what is it?. *American documentation*, 19(1), 3-5.
- [12] Cerqueti, R., & Ausloos, M. (2015). Evidence of Economic Regularities and Disparities of Italian Regions From Aggregated Tax Income Size Data. *Physica A: Statistical Mechanics and its Applications*, 421(1), 187-207.
- [13] Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, 94, 53-64.
- [14] Cinelli, M., Ficcadenti, V., & Riccioni, J. (2019). The interconnectedness of the economic content in the speeches of the US Presidents. *Annals of Operations Research*, doi:10.1007/s10479-019-03372-2.

- [15] Cochran, J. J., Curry, D. J., Radhakrishnan, R., & Pinnell, J. (2014). Political engineering: optimizing a US Presidential candidates platform. *Annals of Operations Research*, 215(1), 63-87.
- [16] Deng, W., & Pato, M. P. (2017). Approaching word length distribution via level spectra. *Physica A: Statistical Mechanics and its Applications*, 481, 167-175.
- [17] Dragulescu, A. A., Poi, A. X., & Dragulescu, M. A. A. (2014). R Package xlsxjars.
- [18] Dragulescu, A. A. (2014). Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files.
- [19] Drożdż, S., Oświecimka, P., Kulig, A., Kwapien, J., Bazarnik, K., Grabska-Gradzińska, I., Rybicki, J., & Stanuszek, M. (2016). Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331, 32-44.
- [20] Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.
- [21] Ficcadenti, V., Cerqueti, R., & Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications*, 123, 127-142.
- [22] Frimpon, M. F. (2013). A multi-criteria decision analytic model to determine the best candidate for executive leadership. *Journal of Politics and Law*, 6(1), 111-127.
- [23] Greenspahn, F. E. (2016). *Hapax legomena in biblical Hebrew: a study of the phenomenon and its treatment since antiquity with special reference to verbal forms (Vol. 74)*. Wipf and Stock Publishers.
- [24] Guns, R., & Rousseau, R. (2009). Real and rational variants of the H-index and the G-index. *Journal of Informetrics*, 3(1), 64-71.
- [25] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569-16572.
- [26] Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85, 741-754.
- [27] Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87-106.
- [28] Iqbal, F., Binsalleeh, H., Fung, B. C., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231, 98-112.

- [29] Ivie, R. L. (1984). Speaking “common sense”; about the Soviet threat: Reagan’s rhetorical stance. *Western Journal of Communication (includes Communication Reports)*, 48(1), 39-50.
- [30] Jewish Virtual Library. <http://www.jewishvirtuallibrary.org/hapax-legomena> [accessed on-line Feb. 2020]
- [31] Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266(1-2), 511-529.
- [32] Joandi, L. (2012). Productivity Measurements Applied to Ten English Prefixes: A comparison of different measures of morphological productivity based on ten prefixes in English.
- [33] Jockers, M. L. (2014). *Text analysis with R for students of literature*. New York: Springer.
- [34] Kahveci, E. & Odabaş, A. (2016). Central banks communication strategy and content analysis of monetary policy statements: The case of Fed, ECB and CBRT. *Procedia-Social and Behavioral Sciences*, 235, 618-629.
- [35] Katta, O. A. (2018). The Influence of Strategic Potential of a Tennis Game on Effort: Understanding the Best Efforts Clause with oktennis.
- [36] Kolakowski, M., & Neale, T. H. (2006). The President’s state of the union message: Frequently asked questions. Congressional Research Service.
- [37] Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: ‘fat tails’ with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4), 525-539.
- [38] Lambiotte, R., Ausloos, M., & Thelwall, M. (2007). Word statistics in Blogs and RSS feeds: Towards empirical universal evidence. *Journal of Informetrics*, 1(4), 277-286.
- [39] Lee, L., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35, 673-702.
- [40] Light, R. (2014). From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses. *Social Currents*, 1, 111-129.
- [41] Lipka, L. (2010). Observational linguistics, neologisms, entrenchment, and the Tea Party Movement. *Brno studies in English*, 36(1), 96-101.
- [42] Machiavelli, N. (2008). *The Prince*. (J. B. Atkinson trans.). Hackett Publishing.

- [43] Mandelbrot, B. (1966). Information theory and psycholinguistics: a theory of words frequencies. In: P. Lazafeld, N. Henry (Eds.), *Readings in Mathematical Social Science*, MIT Press, Cambridge, MA.
- [44] Manning, C., Raghavan, P., & Schütze, H., (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100-103.
- [45] Metin, S. K. (2018). Feature selection in multiword expression recognition. *Expert Systems with Applications*, 92, 106-123.
- [46] Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.
- [47] Mitkov, R. (2004). *The Oxford handbook of computational linguistics*. Oxford University Press.
- [48] Mohammadi, M. (2016). Parallel Document Identification using Zipfs Law. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora* (pp. 21-25).
- [49] Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3-4), 567-578.
- [50] Morente-Molinera, J. A., Kou, G., Peng, Y., Torres-Albero, C., & Herrera-Viedma, E. (2018). Analysing discussions in social networks using group decision making methods and sentiment analysis. *Information Sciences*, 447, 157-168.
- [51] Morin, E., Hazem, A., Boudin, F., & Clouet, E. L. (2015). LINA: Identifying comparable documents from Wikipedia. *Eighth Workshop on Building and Using Comparable Corpora*.
- [52] Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *J. Open Source Software*, 3(23), 655.
- [53] Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons
- [54] Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5), 323-351.
- [55] Obama, B. (2016). United States health care reform: progress to date and next steps. *The Journal of the American Medical Association* , 316(5), 525-532.
- [56] Ooms, J. (2017). Hunspell: High-performance stemmer, tokenizer, and spell checker for R. R package version 2.3.

- [57] Papadimitriou, C., Karamanos, K., Diakonou, F. K., Constantoudis, V., & Papageorgiou, H. (2010). Entropy analysis of natural language written texts. *Physica A: Statistical Mechanics and its Applications*, 389(16), 3260-3266.
- [58] Park, D., Kim, G. N., & On, B. W. (2016). Understanding the network fundamentals of news sources associated with a specific topic. *Information Sciences*, 372, 32-52.
- [59] Parsons, W. (2001). Scientists and politicians: The need to communicate. *Public Understanding of Science*, 10(3), 303-314.
- [60] Pavlov, A. N., Ebeling, W., Molgedey, L., Ziganshin, A. R., & Anishchenko, V. S. (2001). Scaling features of texts, images and time series. *Physica A: Statistical Mechanics and its Applications*, 300(1-2), 310-324.
- [61] Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112-1130.
- [62] Polizzi, G. (2012). Giacomo Leopardi. In: *Il contributo italiano alla storia del pensiero, ottava Appendice*. Istituto della Enciclopedia Italiana Fondata da Giovanni Treccani, Roma.
- [63] Rayson, P., Archer, D. E., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics conference: CL2007*.
- [64] Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus* 1, 2325.
- [65] Rovenchak, A., & Buk, S. (2018). Part-of-speech sequences in literary text: Evidence from Ukrainian. *Journal of Quantitative Linguistics*, 25(1), 1-21.
- [66] Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112, 10837-10844.
- [67] Savoy, J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics*, 17(2), 123-141.
- [68] Schonhardt-Bailey, C., Yager, E., & Lahlou, S. (2012). Yes, Ronald Reagan's Rhetoric Was UniqueBut Statistically, How Unique? *Presidential Studies Quarterly*, 42, 482-513.
- [69] Schreiber, M. (2010). A new family of old Hirsch index variants. *Journal of Informetrics*, 4(4), 647-651.

- [70] Smith, J. A., & Kelly, C. (2002). Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36(4), 411-430.
- [71] Soules, M. (2015). *Media, persuasion and propaganda*. Edinburgh University Press.
- [72] Takikawa, H., & Sakamoto, T. (2019). The moral-emotional foundations of political discourse: a comparative analysis of the speech records of the US and the Japanese legislatures. *Quality & Quantity*, 1-20.
- [73] Toorawa, S. M. (2012). Hapaxes in the Qurān: identifying and cataloguing lone words (and loanwords). In *New Perspectives on the Qurān* (pp. 215-268). Routledge.
- [74] Wickham H. (2016). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2
- [75] Wickham, H. (2018). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.3.0
- [76] Wallot, S. (2011). *The role of reading fluency, text difficulty and prior knowledge in complex reading tasks*. Doctoral dissertation, University of Cincinnati.
- [77] Yoon, H. G., Kim, H., Kim, C. O., & Song, M. (2016). Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. *Journal of Informetrics*, 10(2), 634-644.
- [78] Yuan, H., Lau, R. Y., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, 91, 67-76.
- [79] Zipf, G. K. (1935). *The psycho-biology of language*. Oxford, England: Houghton, Mifflin.
- [80] Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Cambridge MA. Reading: Addison-Wesley.