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

Valerio Ficcadenti[1,4,*] Roy Cerqueti[1,4] Marcel Ausloos[2,5]
Gurjeet Dhesi[3]

[1] Department of Economics and Law
University of Macerata
Via Crescimbeni, 20 - 62100 Macerata, Italy
roy.cerqueti@unimc.it
ficcadentivalerio@gmail.com

[2] School of Business
University of Leicester
University Road. Leicester, LE1 7RH, UK
ma683@le.ac.uk

[3] School of Business
London South Bank University
London SE1 0AA, UK
dhesig@lsbu.ac.uk

[4] School of Business
London South Bank University
London SE1 0AA, UK

[5] 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

*Corresponding author.

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 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 [37], where the authors study the regularities of words occurred in blogs and [30], where the authors propose a model for assessing borrowers' defaults on loans by analyzing texts on the available descriptions of such loans. Chan and Chong (2017) pay peculiar attention to the exploration of the financial texts for their relevant informative content (see [12]). In the same context, Yuan et al. (2016) discuss the determinants of the success of crowdfunding by following a text analysis approach (see [74]).

Nowadays, politicians use social networks to inform their voters, therefore they calibrate the messages on the bases of objectives to be addressed (see [73] 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 study of the communication of US Presidents (and also candidates to presidency or Presidents' media importance [56]) and its socio-economic relevance (see e.g. [14, 21]).

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. At this aim, we study the collection of the hapaxes in each speech.

A detailed discussion of the scientific motivations behind the paper is carried out separately, in the next 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 [53] 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 below – has been attained through data mining techniques (for a survey on text mining, see e.g. [19, 44]). After such a treatment phase, we have obtained 951 speeches to be investigated. The collection of the hapaxes of the speeches has been stored in 951 speech-based sets; such sets have been merged together. The resulting merged set contains all the words that have been hapaxes at least in one speech, along with their frequencies in the overall set of speeches. To clarify this conceptualization of the frequency, think that if the frequency of a hapax is 5, then such a word has been a hapax in five different speeches. This said, in the overall corpus, the maximum hypothetical frequency of a word is 951; for a given President, the maximum hypothetical frequency of a hapax word is equal to the number of speeches he gave (or was recorded, rather) whilst the minimum one is 1.

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 [52]. In according with other linguistic studies [47, 58, 1, 2, 8, 61], we have tested the validity of the Zipf-Mandelbrot law in properly fitting the data by implementing a best-fit optimization procedure [75, 76, 41]. In this preliminary phase, as we will see, 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 by Ausloos (2013) in the science measurement context of the scientists' coauthors (see [5]). Such an indicator is a replication of the H-index – where H stands for Hirsch, who invented it in [24] – used to evaluate scientific research. It is crucial to recall that the value of the H-index of a scientist is $\bar{H} \in \mathbb{N}$ when $\bar{H}$ is the maximum number of papers authored by the scientist which have been cited at least $\bar{H}$ times (see [23] for a more detailed description and [65] 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 of 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 [36], 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 [11]. An interpretation of such an effect will be also presented.

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 [48], 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 [20], which is quite close to our study. 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 [20] 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 [20] 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. Indeed, hapaxes represent a very specific subset of the tokens which compose the corpus of the speeches, and so they have to be extracted from the original dataset by implementing a specific procedure; secondly, the scientific ground of [20] 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, 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. To do so they pickup specific words and repeat those a predefined number of times. Therefore, a global analysis of the speeches structure, like the one presented in the quoted paper, is suitable to address the general framework of the messages delivered, modeling the range of the frequencies. Differently, 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.

The rest of the paper is organized as follows. Section 2 contains a wide discussion on the scientific motivations behind the paper; such a sections includes also a subsection with the motivations behind the investigation methodology. 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.

Supplementary material is devoted to a discussion of some salient hapaxes under a linguistic perspective, in the light of supporting the relevant role of the words pronounced only once in the overall US Presidents decisions on communication strategy.

2 Motivations of the paper

This section contains some arguments supporting the research and scientifically motivating the worthiness of the proposed study.

The exploration of the hapaxes goes much further than usual text content or structure analyses; hapaxes have a special meaning (see e.g. [6, 55, 18, 4, 15, 43]). 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 [59].

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 [27, 68], 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 Joandi (2012), 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 [31] and 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 [40]. 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, Holmes (1994) 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 (see [26]). In Smith and Kelly (2002), the authors are pointing to the importance of the hapaxes within a corpus to extract information about the writers stylistic changes (see [66]).

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 [67]). 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 [63] 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.

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. The most part of them is associated to some peculiar characteristics and leads to crucial conjectures:

(i.1) first, 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]):

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

(i.2) second, 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;

(i.3) third, 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;

(i.4) fourth, 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.

While property (i.1) is supported by a formal definition in the linguistic literature, characteristics (i.2), (i.3) and (i.4) are proposed as conjectures. A formal linguistic study would be needed to

state definitively if such conjectures are true. Unfortunately, there is no room in this paper to face the problems raised in $(i.2)$, $(i.3)$ and $(i.4)$ (see some comments on these aspects in Subsection 5.1). However, it is clear that such conjectures contribute to motivate the worthiness of the exploration of the hapaxes of the official speeches of the US Presidents.

The analysis of the frequencies of the hapaxes over a large set of speeches provides additional insights. Indeed, all the points listed above should be interpreted in the more complex environment of several Presidents and occasions.
In this respect, we provide some (further) research question considerations:

($ii.1$) Largely pronounced hapaxes raise a question about the presence of recurrent themes, treated by the President(s) in the light of the motivations and conjectures $(i.1) - (i.4)$ listed above.

($ii.2$) 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.

($ii.3$) The employment of specific hapaxes by several Presidents raises a question about the existence of a sort of linguistic code.

($ii.4$) Since hapaxes are rather rare words, one can conjecture that they are taken from a dictionary black box, only containing the never used words.

Notice that also $(ii.1)$ $(ii.2)$, $(ii.3)$ and $(ii.4)$ represent research questions, to be interestingly explored in a detailed way, but in future research (some comments on these aspects are contained in Subsection 5.1). However, as already stated for cases $(i.1)$, $(i.2)$, $(i.3)$ and $(i.4)$, these conditions $(ii.)$'s can be seen as a relevant ground for proceeding in the analysis of the hapaxes in the considered corpus of the US Presidents' speeches.

To sum up, we point out that the analysis of the hapaxes here carried out open a wide number of research questions, and that's exactly the philosophical target of our research activity.

We notice that the linguistic perspective has led to many ancient English uses like "goodby"[1] or "pursueth"[2]. Such occurrences are left in the list as they are. Indeed, the method here presented works in order to detect hapaxes even if they consist in old English terms. The technique is not sensible to linguistic changes because a word employed just once in a speech will result in the hapaxes list and, just hypothetically, could be found in what we call 'core of hapaxes'. Consequently, the presented procedure allows researchers to investigate specific hapaxes in order to explore their usage in the past. In this sense, it allows also a proper contextualization of the birth of neologisms,

[1] https://en.wiktionary.org/wiki/goodby

[2] https://en.wiktionary.org/wiki/pursueth

hence leading to the understanding of the socio-political events that generate the needs of certain terminology (e.g. the use of 'economic refugee', as described in [69]).

2.1 Motivations for the employed methodology

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 [46, 49], 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,... is an integer allowing some hierarchical ranking of the CAs; $r = 1$ being the most prolific coauthor with 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 [24, 25, 60] "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,

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

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 [38].

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 H) or, conversely, such an imitative behaviour does not take place.

Finally, to further support the worthiness of our proposal, we can say that 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 [22] 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.

3 Data

This section is dedicated to describe the data collection process. It is realized thanks to an R routine implemented by the means of different packages: *"xml2","rvest", "stringr", "xlsxjars", "xlsx"* along with their respective dependencies (for further packages' information: [34, 70, 71, 16, 17]).

The procedure is close to the one presented in [20, 13]. Indeed, the data collection process is the same, in fact it is just summarized here, while the new steps to get the hapaxes are fully detailed.

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

Notice also that 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.

The web scraping routine phase is implemented on the Miller Center web site, specifically on the page where the speeches are dynamically showed [3] (see [32, 51] for web scraping applications when R is used). A series of exceptions of technical nature have been corrected during this phase.

1. The downloaded texts contain many typos. Therefore, the second phase is devoted to taking care of them in an automatic way. Errors like: *"you.Therefore"*, *"10,000of"* or *"thePresident"* impede an appropriate tokenization, therefore specific rules are created thanks to the regular expressions, in this way the flaws can be easily corrected. (for a formal definition of the recular expressions, see e.g. [45])

2. Further issues arise from the interactions of the Presidents with the audience or with the journalists. It could happen that the speakers are interrupted by applause, laughters, singed slogans or very loud screams at which the President could sometimes respond; or when the Presidents are dealing with press conferences, they could receive questions from different newspapers and they have to respond. These elements are reported into the speeches' transcripts. We consider them as noise because our interest concerns the Presidents' words stated without any external conditions that influence the speakers' flow planned before the public talk is started. For this reason all the transcripts' segments that come after the first journalists' question are excluded. Thus, by employing the regular expressions again, the audience interferences and all the words that come after the first journalists' intervention have been eliminated.

   Another well established types of Presidential public meeting are the debates which are characterized by a dialogue among candidates and/or journalists. These transcripts have peculiar complexity, whence have been removed. In particular, we cancel 13 Debates and 1 Conversation from the original Miller Center database.

3. Another important aspect is the length of the speeches. The first column of Table 1 reports the number of words per speech that results at this phase of the procedure which is dedicated to the control of the speeches length. We have excluded the speeches with less then 132 words, for three main reasons. First, the number of hapaxes in a text is positively correlated with its

[3] `https://millercenter.org/the-presidency/presidential-speeches`

length, so very short texts could lead to a unreasonably high relative frequency of hapaxes. Second, even if any outliers analysis of the numbers in column 1 of Table 1 would lead to a truncation at the first three speeches, because of the relevant difference between the third and the fourth speech, we decided to cut off the first five ones. Indeed, four out of the first five shortest speeches are Press Conferences, including the fifth shortest one. Press conferences contain the highest level of 'impurities' due to questions from journalists, and this represents a bias for us. To remove such an inconsistency, press conferences have been treated by applying a specific rule to detect the questions from the audience. We remove them, along with the related responses. This cleaning phase, together with the previous one led to transcripts containing only the statements that Presidents wanted to deliver before any question from anyone. This said, the unique talk of the first five listed in Table 1 that is not a press conference is the 'Argument before the Supreme Court in the Case of United States *vs*. Cinque' that has been removed from the analysis for its insufficient length. At this point, the remaining speeches' transcripts amount to 951.

4. At this stage, it is necessary to treat the residual typos. We do it here because after the reduction of the number of speeches, the potential errors to be controlled is lower. We check the text by using the *"hunspell"* R package [54], with the English dictionary. We extract and correct all the spelling errors within each talk. But a list of 7716 exceptional tokens that are not found in the English dictionary results. 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 spelled in the past like: "regrassing" or "tofore". These types of exceptions enter into the list of potential typos but they cannot be considered errors to be modified because we assume that the Presidents have pronounced them into a specific context that require such uses. Therefore, in order to make the distinction between the exceptions just described and the flaws that have to be modified, one looks for all the list 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 research' results. 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 findings [4]. The potential flaws with highest degree of uncertainty are adjusted or not on the bases of the majority criterion. It means that one uses the Google's exact search and then adopts the most common phrases within the first 10 results hereby found. Thank to this correction process, 3851 improvements are applied. Consequently, the remaining locutions cannot be considered wrong. This step is the only one requiring the human judgment; therefore, it is the only reason to define the nature of the procedure as *semi-automatic*.

5. Finally the tokenization phase (for a formal definition and some practical examples see [42]) is realized. Here the R library *"tokenizers"* [50] has been used to divide the text thanks to the blank space as a separator; the punctuation is not considered as a separator except for the apostrophes. As an outcome, we obtain a vector for each speech whose components are the tokens.

6. The tokens that occur just once per speech are selected and saved as a list. From this resultant group, a table of frequencies is built by counting the hapaxes' occurrences into each corpus used. The most common hapaxes are reported in Table 2.

The whole resulting list 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 3, column (a), which allows to have a view of the frequencies' collection properties.

Despite of 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 basic algorithms and human readings in order to reach an expectedly very low set

[4] An example of this type is 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 in a systematic way and for being coherent with the method adopted, the exact search is run. In the contest of the example, the research returns many other sources where the words' correct forms are adopted as expected.

N. words	Date and Title	Presidents
5	December 31, 1966: Press Conference	Lyndon B. Johnson
6	August 18, 1967: Press Conference	Lyndon B. Johnson
13	November 17, 1967: Press Conference	Lyndon B. Johnson
83	February 24, 1841: Argument before the Supreme Court in the Case of United States v. Cinque	John Quincy Adams
94	July 31, 1991: Press Conference with Mikhail Gorbachev	George H. W. Bush
132	March 4, 1793: Second Inaugural Address	George Washington
139	December 11, 1941: Message to Congress Requesting War Declarations with Germany and Italy	Franklin D. Roosevelt
143	June 22, 1877: Prohibition of Federal Employees' Political Involvement	Rutherford B. Hayes
152	February 11, 1861: Farewell Address	Abraham Lincoln
156	April 5, 1792: Veto Message on Congressional Redistricting	George Washington

Table 1: The table contains the 10 shortest transcripts at the phase of the procedure dedicated to the control of the speeches length, described in Section 3.

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 misspelled 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 still remains the possibility of leaving errors due to natural human predisposition to commit error in a manual control – operational risk. In fact, 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 with respect to [20]. 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 to conclude 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 see below, the procedure to determine the analyze the hapaxes and identify their 'core' combines the H-index (see [24]) 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.

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

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 2: The most frequent 61 hapaxes, along with their frequencies.

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 3: 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.

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 [13, 20]. 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 [63], the authors have used 245 US Presidents' speeches.
- [62] 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 [39] 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 [64] focuses only on the Ronald Reagan's discourses.
- In [33], 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 a digitalization of old texts sometimes prevents the employment of a complete corpus, mainly for the lack of access to the non-digital texts or for the lack of scanning systems and devices like the OCR (Optical Character Recognition). 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).

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:

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

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 4, and Figure 1 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 [24] 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.

By employing $\bar{H}$ and the best-fit curve defined in (4), with parameters in Table 4, block (a) –

		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
(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 4: (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.

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 $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{R}}$, 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

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

while the relative measure is

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

5 Results and discussion

The results of the best-fit exercise are reported in Table 4, 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 1 further supports such a result by proposing a visual inspection of the fit. A mere power (Zipf) law gives a much worse fit.

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.

Thus, by applying formulas (5) and (6), and by using the values listed in Table 4, section (a),

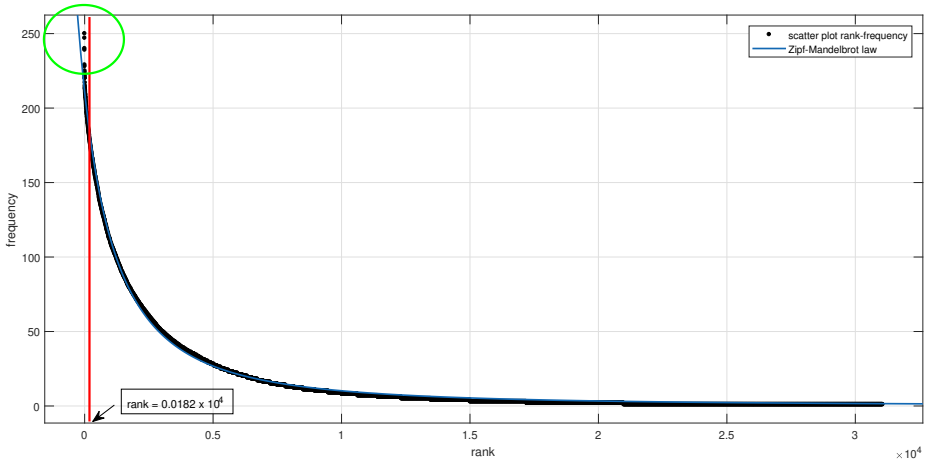

Figure 1: Best-fit curve, according to equation (4) and calibrated parameters in Table 4, rows (a). The scatter plot of the original sample is juxtaposed for a better comparison; the agreement is very good; data and fits are hardly distinguishable from each other. Notice the slight deviations at low ranks (green circle in the Figure), suggesting the presence of king and vice-roy effects (see e.g. [11]). The red vertical line points to $\bar{H} = 182$, which delimitates the core of the hapaxes.

a straightforward computation gives that

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

and

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

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 3 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 2 and Table 4, rows (b) for the details. The 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 2). Such deviations are confirmed also by the wider confidence

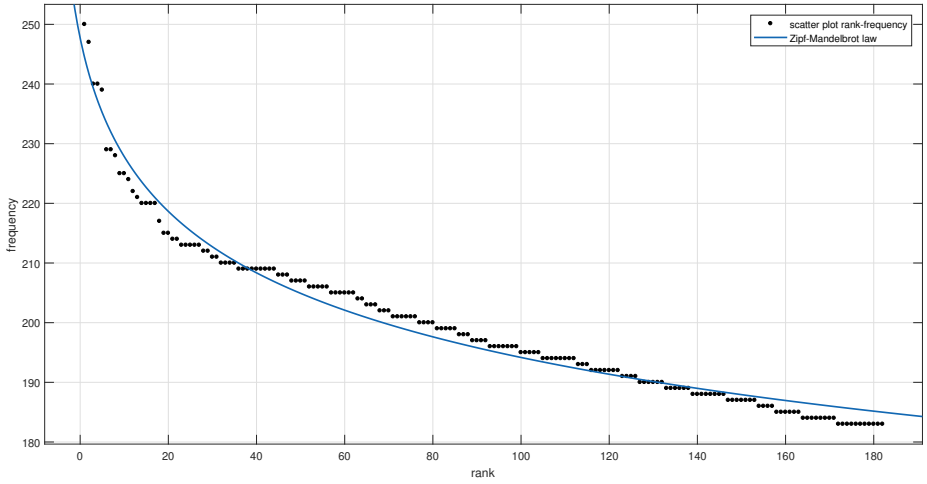

Figure 2: Best-fit curve, according to Eq. (4) and calibrated parameters from Table 4, rows (b) for the case of the hapaxes in the core. The scatter plot of the original sample of the core is also shown for comparison purposes; the agreement is visually good.

intervals resulting in the case of the core with respect to those coming from the overall sample, see Table 4 rows (a) and (b). 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 1 and 3). 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. [36, 5, 11]). 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 4, 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 some very relevant hapaxes (those in the core), whose meaningfulness has to be discussed, also in the light of the role of the hapaxes in a speech (see Section 2 on this). Therefore, a discussion of some of the most frequent hapaxes resulting from the analysis is reported here to show the potentiality of the presented approach. We have also implemented a specific poly-grams analysis to explore the context in which the hapaxes occur.

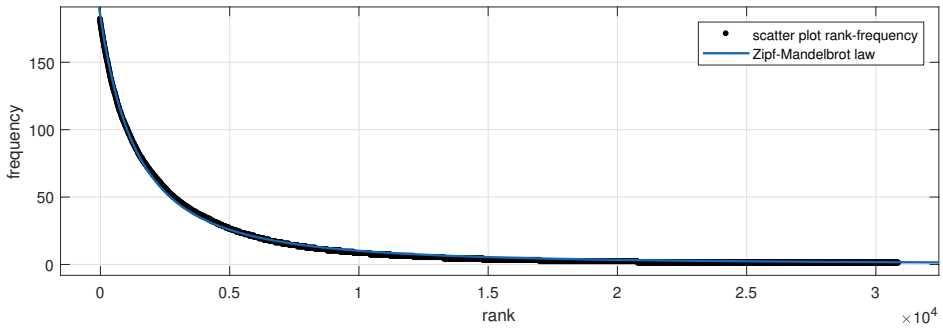

Figure 3: Best-fit curve, according to Eq. (4) and calibrated parameters in Table 4, block (c), for the case of the hapaxes excluding the core. The scatter plot and the fitted curve are not distinguishable. The deviations at the low rank shown in Figure 1 do not appear, thus leading to the statement of the presence of king and vice-roys effects for the elements of the core in the respect of the overall sample.

Furthermore, still in the line of highlighting the implication of our approach, we report here also a discussion of some hapaxes appearing strictly once in the entire corpus.

Let us start checking which Presidents have used the most common hapax *sense* (see Table 5, 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 arguments[5], 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

[5] www.u-s-history.com/pages/h735.html

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 "America hates war. America hopes for peace. Therefore, America actively engages in the search for peace." '*. 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 5 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 5). 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). In [35] it is reported that 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 "President's 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 restored[6]. Furthermore, thank to a visual inspection

[6] `http://www.whitehousemuseum.org/special/renovation-1948.htm`

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, I'm 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 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 a computational and 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, 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 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 Conclusions and future research

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 talks 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 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 and the consequent analysis.

From Section 5 the ability of our method in responding to the research questions clearly emerges. Indeed, the hapaxes legomenon occurring in a collection of texts can be huge and their meanings are remarkable, as already stressed above. One should go trough 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 spurious results and 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.

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. [57, 3, 4, 18].

The analysis of how such speeches should be received has been already a little bit discussed in [2, 4, 72]. 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 2). 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 of 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 Ausloos and Cerqueti (2016), where a rank-size law was also discovered (see [7]). 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.

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 5: 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] 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.

[12] Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. Decision Support Systems, 94, 53-64.

[13] 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.

[14] 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.

[15] Deng, W., & Pato, M. P. (2017). Approaching word length distribution via level spectra. Physica A: Statistical Mechanics and its Applications, 481, 167-175.

[16] Dragulescu, A. A., Poi, A. X., & Dragulescu, M. A. A. (2014). R Package 'xlsxjars'.

[17] Dragulescu, A. A. (2014). Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files.

[18] Drożdż, S., Oświecimka, P., Kulig, A., Kwapień, 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.

[19] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

[20] 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.

[21] 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.

[22] 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.

[23] Guns, R., & Rousseau, R. (2009). Real and rational variants of the H-index and the G-index. Journal of Informetrics, 3(1), 64-71.

[24] 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.

[25] 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.

[26] Holmes, D. I. (1994). Authorship attribution. Computers and the Humanities, 28(2), 87-106.

[27] http://www.jewishvirtuallibrary.org/hapax-legomena

[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] 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.

[31] Joandi, L. (2012). Productivity Measurements Applied to Ten English Prefixes: A comparison of different measures of morphological productivity based on ten prefixes in English.

[32] Jockers, M. L. (2014). Text analysis with R for students of literature. New York: Springer.

[33] 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.

[34] Katta, O. A. (2018). The Influence of Strategic Potential of a Tennis Game on Effort: Understanding the Best Efforts Clause with oktennis.

[35] Kolakowski, M., & Neale, T. H. (2006). The President's state of the union message: Frequently asked questions. Congressional Research Service.

[36] 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.

[37] 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.

[38] Lee, L., Bozeman, B. (2005). The impact of research collaboration on scientific productivity. Social Studies of Science, 35, 673-702.

[39] 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.

[40] Lipka, L. (2010). Observational linguistics, neologisms, entrenchment, and the Tea Party Movement. Brno studies in English, 36(1), 96-101.

[41] 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.

[42] Manning, C., Raghavan, P., & Schütze, H., (2010). Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103.

[43] Metin, S. K. (2018). Feature selection in multiword expression recognition. Expert Systems with Applications, 92, 106-123.

[44] 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.

[45] Mitkov, R. (2004). The Oxford handbook of computational linguistics. Oxford University Press.

[46] Mohammadi, M. (2016). Parallel Document Identification using Zipf's Law. In Proceedings of the Ninth Workshop on Building and Using Comparable Corpora (pp. 21-25).

[47] Montemurro, M. A. (2001). Beyond the Zipf-–Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications, 300(3-4), 567-578.

[48] 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.

[49] Morin, E., Hazem, A., Boudin, F., & Clouet, E. L. (2015). LINA: Identifying comparable documents from Wikipedia. Eighth Workshop on Building and Using Comparable Corpora.

[50] 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.

[51] Munzert, S., Rubba, C., Meiβner, P., Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons

[52] Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary physics, 46(5), 323-351.

[53] 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.

[54] Ooms, J. (2017). Hunspell: High-performance stemmer, tokenizer, and spell checker for R. R package version 2.3.

[55] Papadimitriou, C., Karamanos, K., Diakonos, 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.

[56] 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.

[57] 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.

[58] 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.

[59] Polizzi, G. (2012). Giacomo Leopardi. In: Il contributo italiano alla storia del pensiero, ottava Appendice. Istituto della Enciclopedia Italiana Fondata da Giovanni Treccani, Roma.

[60] Rousseau, R. (2006). New developments related to the Hirsch index. Science Focus 1, 23–25.

[61] Rovenchak, A., Buk, S. (2018). Part-of-speech sequences in literary text: Evidence from Ukrainian. Journal of Quantitative Linguistics, 25(1), 1-21.

[62] 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.

[63] Savoy, J. (2010). Lexical analysis of US political speeches. Journal of Quantitative Linguistics, 17(2), 123-141.

[64] Schonhardt-Bailey, C., Yager, E., & Lahlou, S. (2012). Yes, Ronald Reagan's Rhetoric Was Unique—But Statistically, How Unique? Presidential Studies Quarterly, 42, 482-513.

[65] Schreiber, M. (2010). A new family of old Hirsch index variants. Journal of Informetrics, 4(4), 647-651.

[66] Smith, J. A., and 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.

[67] Soules, M. (2015). Media, persuasion and propaganda. Edinburgh University Press.

[68] 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.

[69] Van Dijk, T. A. (1997). What is political discourse analysis. Belgian Journal of Linguistics, 11(1), 11-52.

[70] Wickham H. (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2

[71] Wickham, H. (2018). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.3.0

[72] Wallot, S. (2011). The role of reading fluency, text difficulty and prior knowledge in complex reading tasks. Doctoral dissertation, University of Cincinnati.

[73] 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.

[74] Yuan, H., Lau, R. Y., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. Decision Support Systems, 91, 67-76.

[75] Zipf, G. K. (1935). The psycho-biology of language. Oxford, England: Houghton, Mifflin.

[76] Zipf, G. K. (1949). Human behaviour and the principle of least-effort. Cambridge MA. Reading: Addison-Wesley.