The impact of lacking metadata and data truncation for the measurement of cultural and linguistic change using the Google Ngram datasets

Alexander Koplenig[1]

1 Institute for the German language (IDS), Mannheim, Germany.

Abstract

Background: The Google Ngram Corpora (GNgC) seem to offer a unique opportunity to study linguistic and cultural change in quantitative terms. To avoid breaking any copyright laws, the datasets are not accompanied by any metadata regarding the texts the corpora consist of and the data are truncated to prevent an indirect conclusion from the n-gram to the author of the text. Some of the consequences of this strategy are analyzed in this paper.

Data: The complete 1-gram and 2-gram datasets Version 2 (July 2012) for each of the following languages are used for the data analysis: British English, French, German, Italian and Spanish. In addition, data from the British National Corpus (BNC) and the German Reference Corpus (DeReKo) was analyzed in order to study the effects of data truncation.

Methods and Results: (1) By re-analyzing the example of measuring censorship in Nazi Germany, which received some widespread attention and was published in a *Science* paper that accompanied the release of the GNgC data, I show that without proper metadata, it is far from clear whether the results actually reflect any kind of censorship at all. (2) Using maximum likelihood to estimate the exponent of Zipf's law for all investigated languages based on both the 1-gram and the 2-gram GNgC data, as well as the BNC and the DeReKo data, the impact created by the truncation of the data is shown. A method to circumvent the problems is demonstrated.

Conclusions: The findings imply that because of lacking metadata and data truncation, conclusions based on the GNC have to be treated very cautiously. Instead of speaking about general linguistic or cultural change, it seems to be preferable to explicitly restrict the results to linguistic or cultural change *as it is represented in the Google Ngram data.*

Introduction

When the Culturomics team, in collaboration with Google, made its huge Google Ngram Corpus (GNgC) available for public use [1], many researchers (including the author of this paper) hoped that this vast amount of data would enable them to study linguistic and cultural change with unprecedented accuracy as it contains roughly 4% in the 2009 version [2] and even 6% in the 2012 version of all books ever published [3]. Since then, quite a few papers presenting more or less innovative approaches to the measurement of linguistic and cultural change have been published [2,4–18]. At the same time, several comments by various linguists, published mostly in blogs, strongly criticize the results based on the GNgC [19–26].

In this paper, I first want to discuss and expand on one fundamental objection raised against the GNgC data – the problem of lacking metadata – and demonstrate empirically why this is a serious methodological problem that could bias conclusions based on the GNgC data or even render them invalid. Secondly, I will focus on the problem that the GNgC only contains n-grams that occur at least 40 times in the corpus as a whole (n-grams are sequences of word-strings. A 1-gram is a sequence of one word, such as "pressure", while a 2-gram is a sequence of two consecutive words, such as "the pressure". So, "drop the pressure" would be a 3-gram in this terminology). While this fact is explicitly mentioned by the GNgC team and – as the lacking metadata – has to do with legal reasons [2], I believe that this fact has not received the attention it actually deserves due to the nature of word frequency distributions in general [27].

Cultural and linguistic change on the micro scale: the importance of meta-information

Frequency as the number of times a certain event has occurred in a certain time span is clearly one of the most important proxies used both in the empirical sciences and the digital humanities. Broadly speaking, if one phenomenon is more frequent than another, we assume that this has "some kind of significance" as [28] puts it. In epidemics for example, researchers measure the frequency of polio cases per year to assess the effectiveness of certain polio vaccines. Applied to the measurement of

cultural and linguistic change, this means that empirically observable changes in the frequency of different cultural events approximate cultural and linguistic change. For example, if we detect that, compared to previous decades, less and less hats are being sold, one obvious interpretation is that – all other things being the same – hats are not as fashionable as they once used to be. It seems to be equally reasonable to assume that the attitude towards racial discrimination in a society is changing, if we find out that the frequency of racial slurs in different newspapers has declined. Linguistic change is no exception to this line of reasoning: a linguist who claims that the English language has become less formal in recent decades could operationalize formality as the frequency of passivization [29] and count the (relative) frequency of sentences that are in passive voice at different successive moments in time. It would serve as empirical evidence of the linguist's claim if the resulting series decreases over time. Given – again – that all other things have remained the same, if the relative frequency of passivization declines, the researcher could argue that the investigated language displays an overall trend to become less formal; at least if we restrict our investigation to the analysis of written language record.

The fundamental problem with the GNgC data is that we just cannot be sure if *all other things* have really remained *the same*, because there is no information about the books of which the GNgC data consists. Or put differently, we cannot check if the different diachronic book samples really represent similar things at different moments in time. One could even claim that the representativeness of a corpus is the most fundamental methodological problem in corpus linguistics, or less dramatically, one could call it a *thorny issue* [29,30]. It is argued in [31] that, while this is a principle problem that can hardly be solved in synchronic corpus design, the situation is much more complex in the design of a diachronic corpus for broader multi-purpose research goals. This is due to several reasons. Obviously, fewer texts are available for earlier periods. Furthermore, the corpus compiler has to decide what registers should be included in the corpus. [31, p. 252] give an instructive example: medical research articles in 1700 were more like case studies written in the form of personal letters to the editor of the journal. Those texts are very different from the standard journal article that

would be accepted for publication in the *Lancet* today. Therefore, it would be quite naïve to compile a diachronic corpus based on medical research articles and to argue that the language it attempts to represent has changed in the last three hundred years. What is allegedly a form of language-internal change should rather be considered a language-external conventional change (see [32], for details and [33], for advantages and problems of this distinction). Of course, this does not imply that language-external changes, or changes that reflect register differences, are not interesting phenomena as [32,34] convincingly demonstrated. However it is of utmost importance to have the ability to separate language-external effects from language internal ones, especially in regard to historical linguistic change, as [35] shows. Because of the problems mentioned, the compilation of a corpus will always, at least to some extent, remain subjective. However, using the metadata, that is information about the texts that a corpus compiles, can at least help to assess the subjectiveness of the interpretation made about the types of likely factors that are at play. This lack of metadata is exactly what can and what has been criticized about studies based on the GNgC data [20,21].

It seems fair to point out that the Culturomics team that is behind the GNgC project [2] are very well aware of this problem. According to the FAQs of the accompanying website [1], the project is still attempting to obtain permission to release the full 5.2 million book bibliography, containing information about each book that is in the corpus for each language. It is referenced that the composition of the corpus (at least from 1800-2000) reflects the acquisition strategies of the major libraries that the Google Books project is working with. However, to minimize potential OCR errors [36, p. 5], a filtering method to determine each book's quality filter score is employed. Since only books with a high OCR quality score are included in the corpus, it is an empirical question whether this procedure systematically biases the results.

To ensure that the corpus compilation actually represents the diversity of books published in each year, the books were re-sampled, resulting in a more balanced text collection. This corpus was then published as a separate corpus named English One Million [36, p. 13]. Unfortunately, this corpus only

exists for the English language and - rather surprising - only for the first published version (July 2009), but not for the second version (July 2012).

In the next section, I will show why the concerns raised due to the lacking metadata are far from being exaggerated. On the contrary, one could even defend an extremist position and argue that, even with a perfectly balanced corpus, meaning that we could be completely sure that the books in the GNgC would be a perfect random sample of all books published in each particular year, without proper metadata, one still could not distinguish language-external changes from language-internal ones. The reason is, according to [36, p. 13], that it is very likely that the types of books that are published are changing as a function of time.

Results

Censorship in Nazi Germany - language internal or external?

In their *Science* paper that accompanied the release of the GNgC data, [2] argue that censorship can be detected by measuring changes in frequency, for example, the number of times the name of a person is mentioned. It is argued that a decrease in frequency in the period of the German Nazi regime implies that this person was suppressed during this time. On the other side of this spectrum, if a person was supported by the regime, this should be evident by an increase in the frequency of use of his or her name. This assumption is tempting, but can be contested. An increase or a decrease in frequency could mean several different things that have nothing to do with censorship at all. For example, a name like Thomas Mann, a famous German contemporary writer, could appear less often in the time of the Nazi regime, not because he was suppressed, but because less books focusing on literary topics were written in this time. Again, a lack of metadata is crucial. It does not even seem implausible that the books published in times of major wars are systematically different from books published in times of peace: According to the so called *Wehrgesetz* [37], passed by the German parliament in May 1935, all male citizens aged between 18 to 45 were obligated to perform military duties. This means that less people were available who would otherwise have spent their time

writing and publishing books. If we assume that the books that would have been written by military-aged persons are different from persons who are not liable for military service, then the language samples in those years are systematically biased because of language external reasons. Again, this is an empirical question that can only be answered with the help of metadata.

But even if we could rule out this fact, one could raise a general objection against the assumption that frequency of use approximates suppression, because it is not possible to deduce the contexts in which a name is being used. Consider Emmanuel Goldstein, the infamous public enemy Nr.1 in Orwell's 1984: he is mentioned or appears in almost every public speech by Big Brother. But does this imply that he benefits from government propaganda or does this just mean that he is used for governmental purposes, to produce "anger and fear" as Orwell puts it? Collectively, I believe that these points cast doubt on the assumption of using frequency as the sole indicator of suppression, especially in a corpus that consists exclusively of books and not periodicals such as newspapers.

To understand why contextual information also matters in the example of quantitatively detecting "Nazi repression de novo" [2, p. 4], one first has look at the method developed by the Culturomics team. The authors start with a list of 56,500 people, who, according to the Wikipedia database, are "the 500 most famous individuals born in each year from 1800 – 1913" [36, pp. 40] and first remove individuals with a mean frequency of less than 5×10^{-9} to make the calculations less susceptible to random fluctuations. For each of the remaining 2,976 persons, three mean values are calculated: (1) the mean frequency of the name of a person from 1925-1933, (2) the mean frequency from 1933-1945 and (3) the mean frequency from 1955-1965. Using simple Newtonian linear interpolation, one can use (1) and (2) to compute (4), an expected mean value for the year 1939. This value is then compared to the actual mean value by dividing (4) by (2). Thus, a value above one implies that the respective name appears less often than expected when a linear relationship is assumed. On the other hand, a value below one means that a person is mentioned more often than actually expected. As a side effect, this also means that when a person does not appear at all in the database from 1939-1945, one would have to divide the expected value by zero. A typical strategy in language

processing would be to stipulate a small mean value for (2), see for example [38 pp. 131-137]. Instead of using this approach, [2] gave the resulting suppression index of that person a value of 200 [36, p. 40]. If we look at the data published by [2] as supporting online material [39], this is quite a surprising strategy, given the fact that the highest actually calculated value amounts to 127.98. In addition, of the top 50 most suppressed persons, 28 received this artificial value.

Since it is also questionably whether a linear model is the right kind of approach to capture the nature of the investigated relationship, I used the original 2-gram datasets to plot the frequency (relative per one million word tokens) against time. Figure **1** shows the results for four individuals that belong to the 30 most suppressed individuals (A-D) and four individuals that belong to the 50 most enhanced individuals (E-H) according to the analysis of [2, SOM data]. Konrad Adenauer (A), the first post-war chancellor of West Germany, is the individual with the highest actually calculated index value mentioned above (i = 127.98). Given the trajectory of the curve, this does not really look like suppression, but more like the beginning of his political career after WWII. The same seems to hold true for the famous political theorist Hannah Arendt (B), who receives the second highest actual index value aside from Adenauer (i = 98.04): while it is certainly tempting to believe that Arendt suffered from censorship by the Nazis because of her Jewish origin, I do not think that the data reflect this belief. These doubts become even stronger for (C) and (D). The first one is for Theodor Blank (artificial index value of 200), the first post-war defense minister in West Germany. Blank was a member of the *Wehrmacht*, the armed forces of Nazi Germany, and served as a lieutenant during WWII, so it is more than unclear why he should have suffered from any kind of suppression by the regime [40]. Again the plot does not warrant another conclusion.

The plot for (D) is of Bill Haywood (artificial index value of 200). U.S. citizen William Dudley “Big Bill” Haywood was a prominent founding member and the leader of the Industrial Workers of the World [41]. Why he should be one of the main victims of German censorship does not become clear at all. If we therefore look directly into the raw data, we find out that this name is mentioned once in 1928 in one book and ten times in the year 1930 in two different volumes. Between 1933 and 1945 his name

does not appear in any of the books, while he is mentioned 14 times between 1955 and 1965. If this is enough to accept an (artificial) index value of 200, it should at least be questioned.

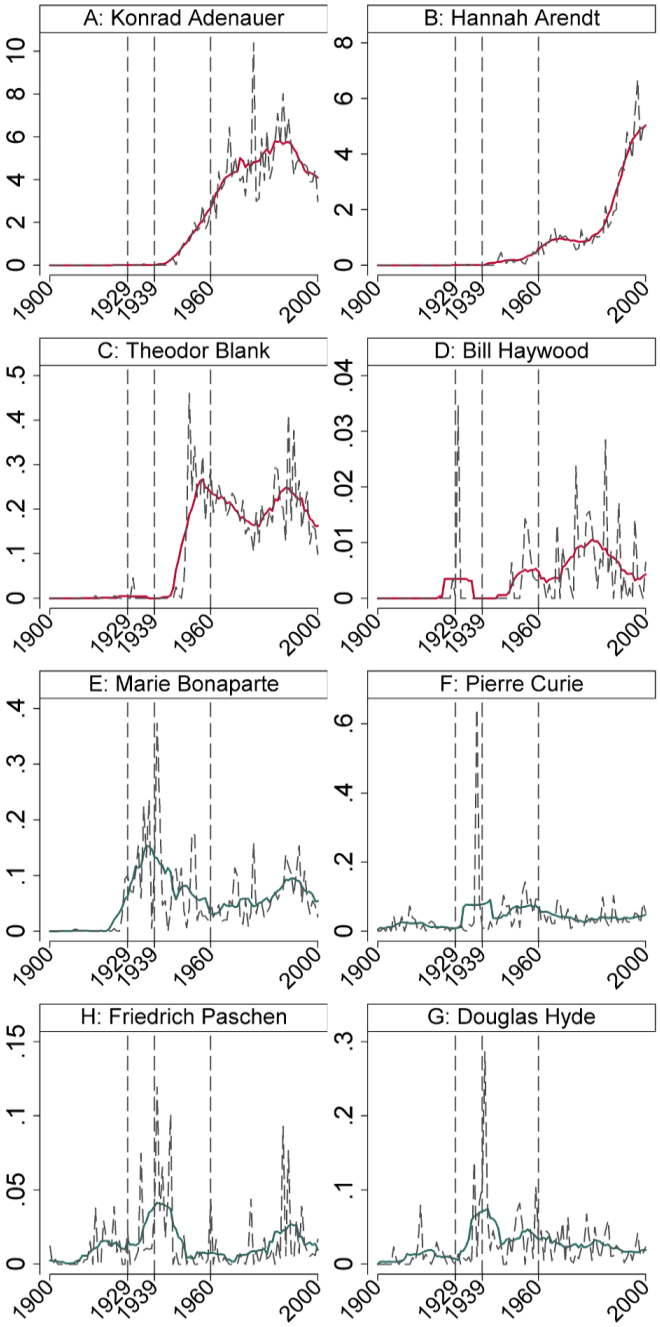

Figure 1: Frequency changes as a function of time (relative per one million 2-gram tokens). (A-D) show the plots of individuals that belong – according to the analysis of [2] – to the 30 most suppressed people in the times of the German Nazi regime. (E-G) show the time series for individuals who belong to the 50 most enhanced persons. The dashed gray lines depict the raw GNgC 2-gram data, while the solid cranberry (A-D) / emerald (E-G) lines add a symmetric eleven-year window moving-average smoother highlighting the central tendency of the series at each point in time.

On the other side of the scale, [36, p. 5] also identified people who benefited from Nazi propaganda. E-H plot the frequency trajectories for four people who belong to the top 50 most enhanced people. (E) is for Maria Bonaparte, the princess of Greece and Denmark. It is possible to interpret the data as if Bonaparte profited from the regime, but one has to ask why this should be the case. Bonaparte worked as an author and as a psychoanalyst and was a close friend of Sigmund Freud [42]. Why the Nazis should enhance her ideas about (female) sexuality [43] is again far from obvious. It seems more likely that the rise in frequency reflects the fact that she published many of her works between 1934 and 1953, and other researchers such as Freud discussed her ideas in their publications. Quite a similar argumentation could be used for the frequency plot of the famous French Nobel laureate Pierre Curie (F). The graph on the bottom left (G) of Figure **1** plots the relative frequency of Friedrich Paschen, who was a physicist like Curie and the president of the renowned Physikalisch-Technische Reichsanstalt (PTR) in Berlin. Again, I believe that the plot reflects the academic career of a scientist and not any kind of enhancement by the Nazi government. On the contrary, his aversion to the regime is illustrated by this story [44, p. 168/169]: when Paschen saw that Nazi supporters had raised a flag with the swastika on the roof of one of the PTR buildings to celebrate the rise to power of the Nazis on 03/05/1933, Paschen ordered the flag to be taken down and brought to his office. Soon after that, he was replaced as the director of the PTR by NSDAP member Johannes Stark in 1933 and forced into "permanent retirement" a few years later. Thus it seems rather unlikely, that his ideas were enhanced in the subsequent period, and classifying him as a beneficiary of the National Socialist regime certainly does not pay homage to his courage.

The last plot (G) tells another interesting story. The rise in frequency of the name of the Irish poet Douglas Hyde does not seem to have anything to do with propaganda enhancement whatsoever, but with the simple fact that he was the first president of Ireland from 1938 to 1945.

One might object that I have only cherry-picked examples where the method developed by [2] did not work particularly well and there are other examples where the method works much better. I encourage any reader to replicate other findings either by using the N-gram viewer that is available

online for public use [1] or – more conveniently – to have a look at SI2, where I prepared time series plots similar to those presented in Figure 1 for both the 50 most suppressed people and the 50 most enhanced persons according to [39]. The dashed gray lines depict the raw GNgC 2-gram data, while the solid cranberry (A-D) / emerald (E-G) lines add a symmetric eleven-year window moving-average smoother highlighting the central tendency of the series at each point in time. I believe that the data presented in this section show that the method is not the best example, when it comes to demonstrating the unprecedented effectiveness and the quantitative precision promised, to gaining new insights into cultural (or linguistic) phenomena [2].

Without access to the list of books that the datasets consist of, it is hardly possible to test any of the assumptions outlined above. However, there seems to be at least some metadata available, to show that the types of books that are included in the GNgC are somewhat different in times of WWII, which could point towards a language-external effect. For the subsequent analysis, I used the total-counts files for each investigated language (cf. Material and Methods). To calculate the average book-length for each year, I simply divided the total number of words that the corpus in each year contains of by the total number of books that are in the corpus in each year.

Figure **2** shows that for all investigated languages (except the Spanish data) the average book lengths are much shorter in the period of WWII (1939-1945) and to a lesser extent in the period of WWI (1914-1918) than in previous and subsequent periods. The Spanish GNgC data does not show this pattern, which could be explained by the fact that South America and Spain did not participate directly in WWII. This result seems to point towards an interesting language-external effect. It certainly shifts the burden of proof towards any researchers who claim the existence of language-internal effects in those particular times.

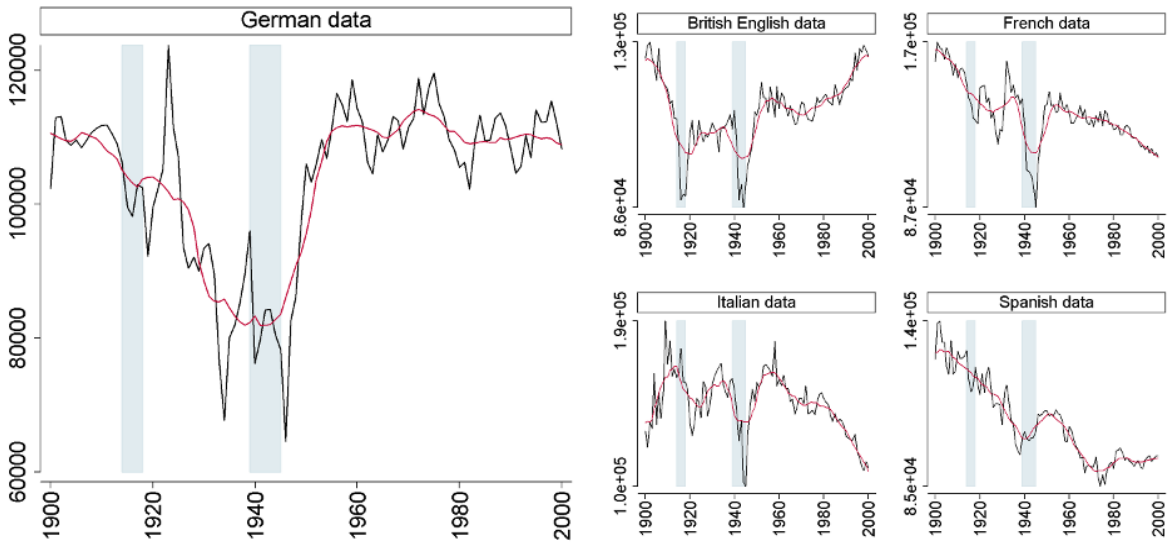

Figure 2: Average book length as a function of time for five different languages. The average book length is calculated by dividing the corpus size in word tokens by the total number of volumes the corpus consists of for each individual year. The black lines depict the raw data, while the cranberry lines add a symmetric eleven-year window moving-average smoother highlighting the central tendency of the series at each point in time. While the series for the different languages show very different overall trends, what stands out is that for all languages, apart from Spanish, the average book length in times of WWII and to a lesser extent in times of WWI was much shorter, pointing towards language-external effects (WWI+WWII times highlighted by the bluish gray shaded parts of the plots).

Data truncation: problematic both on the micro and the macro scale

The diachronic information in the corpus can also be removed by summing up the annual total counts for each n-gram. Again, for legal reasons, the Culturomics team decided to exclude n-grams that appeared less than 40 times across the whole (synchronic) corpus. [36, p. 12] claim that the "most robust historical trends are associated with frequent n-grams". Therefore, removing infrequent n-grams should not alter the results dramatically. Again, I believe that this claim is an empirical question that has to be tested quantitatively. Given the fact that the corpus size for each year strongly increases as a function of time, one has to test whether this arbitrary data truncation imposes a systematic bias on the data, because the vocabulary size of a corpus, i.e. the total number of distinct word types, strongly depends on the size of a corpus, i.e. the total number of word tokens. Unfortunately, as [27,45] show, measures that try to account for this problem all seem to fail.

Using the 1-gram datasets in conjunction with the additional information published in total-counts files (which report the total 1-gram token frequencies for each year including 1-grams that appear less than 40 times), we can calculate the fraction of tokens that are affected by this procedure, since the sum of the 1-gram occurrences in those files includes 1-grams that appear less than 40 times. For example, it only constitutes on average 0.69% (minimum: 0.48%; maximum: 1.39%) of all tokens per year in the British English corpus and 1.35% (minimum: 0.93%; maximum: 2.69%) in the German corpus. This does not seem to be that big of a deal. However, from a corpus linguistic point of view it certainly is: to cut a long story short, for 1-grams only, this approach eliminates approximately 95% of all different 1-gram types, for n-grams where n>1 this figure is even higher [46].

The reason for this quite surprising effect is – of course – that word frequency lists tend to be Zipf-distributed [47,48]. If one assigns rank 1 to the most frequent word, rank 2 to the second most frequent word, and so on, then the frequency *f* of a word and its rank *r* are related as follows:

$$f(r) \propto r^{-\alpha} \qquad (1)$$

In this case, α is a parameter that has to be determined empirically. Thus, Zipf's law explains why only a small fraction of all tokens in the GNgC is eliminated by the truncation approach, because we have a few word types (mostly function words) that account for the overwhelming majority of all tokens.

However, this is only one side of Zipf's law [46]. Following [27, chapter 1.2 & 1.3], one can also model the frequency spectrum *V(m)*, that is the number of word types with frequency *m* in terms of the zeta distribution as:

$$V(m,N) \propto m^{-\alpha} \qquad (2)$$

In the simplest case, α is equal to unity (as Zipf frequently assumed), then [27] shows that:

$$V(m,N) \propto \frac{1}{m(m+1)} \qquad (3)$$

This in turn means that a very large number of word types (mostly content words) occur only very seldom. One direct consequence of (3) is that approximately half of the vocabulary only occurs once in a corpus. Cumulating the frequency spectrum for words that occur less than 40 times means that approximately 95% of all word types are excluded. It is important to note that infrequent words are not only spelling or OCR errors as [36, II.3.] assume, but also words that do not seem unusual to a speaker of a language [46]. It is even more important to keep in mind that Zipf's law can be also extended to n-grams where n > 1, as [49] demonstrate. For example [46] shows that 92% of the trigrams in the Brown corpus (consisting of 500 different English-language texts) only occur once. Since – of course – words do not occur randomly [50], it does not seem appropriate to stipulate *a priori* that truncating the data does not affect the results based on the GNgC data.

The influence of the data truncation of the GNgC can be understood best by using the 2-gram data sets and aggregating the data to calculate 1-gram token frequencies for each word type. For the German data, 66.63% of all word types that occur in the 1-gram data do not occur in the 2-gram data. Table **1** summarizes these results for the different investigated languages.

Language	Number of word types in the GNgC data		Relative loss (in %)	CDF percentiles of the token frequency of excluded word types					
	1-gram	2-gram		p25	p50	p75	p90	p95	p99
British English	6,358,907	2,424,740	61.87	62	93	163	324	538	1825
French	5,544,680	2,177,419	60.73	60	86	138	223	303	588
German	6,527,853	2,178,047	66.63	66	100	171	294	414	926
Italian	3,273,035	1,390,494	57.52	61	88	141	224	301	563
Spanish	4,266,834	1,771,095	58.49	60	85	136	222	306	610

Table 1: The impact of the data truncation. For each language, the total number of word types is calculated for the whole corpus both based on the 1-gram data (column 2) and on the 2-gram data (column 3). Column 4 contains the relative loss which is just the number of word types based on the 1-gram data divided by the number of word types based on the 2-gram data. The remaining columns summarize the (cumulative) distribution of the word token frequency of word types that are excluded due to data truncation in the 2-gram data but are available in the 1-gram data.

A visual inspection of the word types that occur in the 1-gram data but not in the 2-gram data reveals that quite a few of these are either typos, OCR errors or wrongly categorized POS types. However, many word types (mostly content words) that are excluded do not belong to one of those groups and, nonetheless, as [2] themselves point out: the GNgC data is just too big for thorough manual

inspections of this kind. Therefore, to show that the bias resulting from the data truncation is systematic, I calculated the fraction of 1-grams that appear only once (called *hapax legomena*) in a particular year. Is is important to note that this is not a contradiction of what is said above, because even though unigram word types that appear less than 40 times are excluded, word types that appear more often across the whole corpus can – of course – only appear once in one individual year. Figure 3 shows a graph in which both the corpus size and the relative frequency of *hapax legomena* (running from minimum to maximum) are plotted against time. A visual inspection of the line plot confirms what the cross-correlograms for the first differences of both variables (accounting for non-stationary, c.f. Materials & Methods for further details) on the right-hand side of each graph also show: the peak at lag 0 and the negative correlation indicate that as corpus size increases, the fraction of *hapax legomena* decreases for all investigated languages, especially for the French data (r_{GB} = -0.51; r_{FRE} = -0.75; r_{GER} = -0.51; r_{ITA} = -0.47; r_{SPA} = -0.36). Restricting the analysis to the period from 1900 to 2000 results in an even stronger negative correlation at lag 0 for all languages but French (r_{GB} = -0.65; r_{FRE} = -0.73; r_{GER} = -0.82; r_{ITA} = -0.69; r_{SPA} = -0.63).

For inter-language comparisons, the relationship between corpus size and the fraction of *hapax legomena* could be problematic, because (a) the correlation strength and (b) the sizes of the base corpora vary from language to language [3].

Certainly, when it comes to measuring linguistic change, the frequency of occurrence plays a key role. As [10,51] shows, frequent structures undergo different forms of change than infrequent structures, mainly because frequently used structures, i.e. words and phrases, are highly entrenched. Thus, without the possibility of disentangling the different effects that are at work for infrequent and frequent linguistic structures, one has to be cautious when drawing conclusions about general linguistic change.

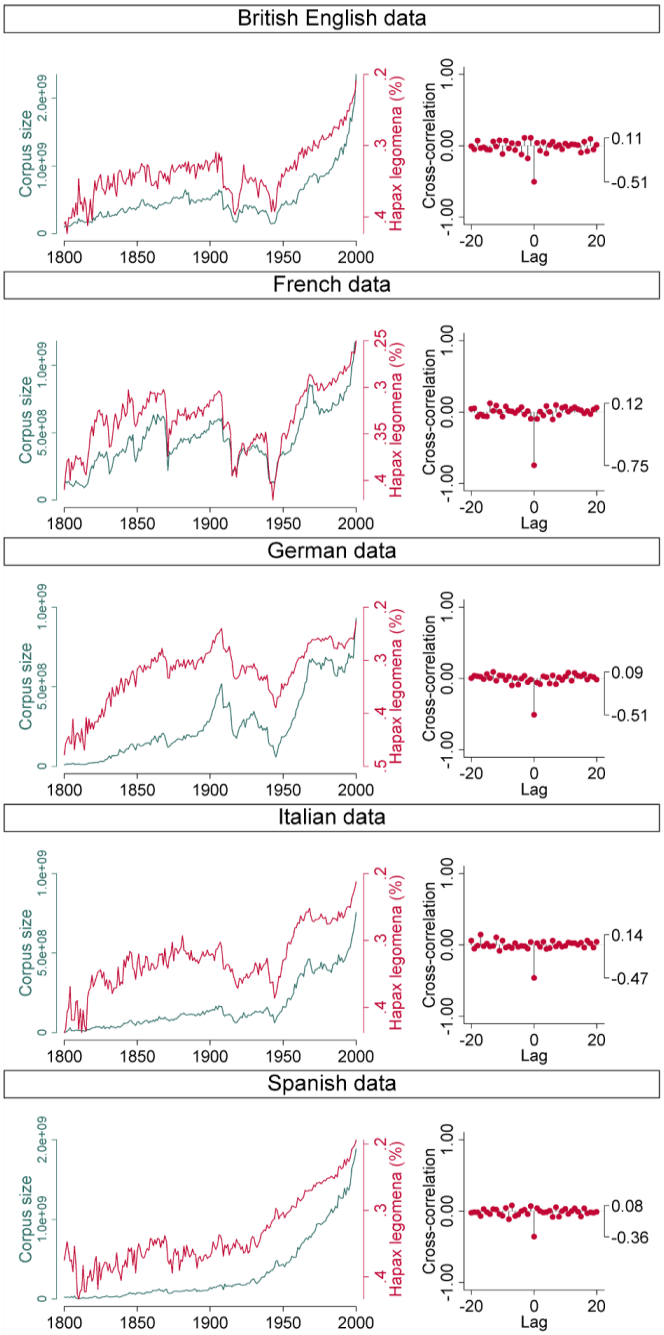

Figure 3: The relationship between the corpus size and the relative frequency of *hapax legomena* (word types that occur only once in the corpus). *Left*: the corpus size (emerald lines) and the frequency of *hapax legomena* (cranberry lines, running from maximum to minimum) as a function of time. *Right*: cross-correlograms for the first differences of both time series (accounting for non-statrionarity). The peak at lag 0 and the negative correlation indicate that as corpus size increases, the fraction of *hapax legomena* decreases for all investigated languages.

The evolution of the Zipf exponent

In this section, I use the evolution of the Zipf exponent as an example to show how the problem raised in the last section can be circumvented, at least for some research questions. For language ontogeny, [52] show that the exponent of Zipf's law tends to decrease over time. They argue that this could be a simple indicator of linguistic or syntactic complexity. In a similar vein, [53] observe a change of the parameters of the Zipf-Mandelbrot law across parallel texts of different time periods and argue that this change is related to changing grammatical encoding strategies.

To show how the problem of data truncation can be tackled, I used a frequency list for British English compiled by [54] of the written part of the British National Corpus (BNC), a roughly 90-million word collection and a frequency list for German of the German Reference Corpus DeReKo (Version 2011) that consists of roughly 4.5 billion word tokens (see [55] for details on the corpus design and compilation).

Both the BNC corpus and the DeReKo corpus are divided into independent sub corpora that vary in size (50% / 25% / 12.50 % / 6.25% / 3.125%) by performing a binomial split (as suggested by [56]). For each corpus, the Zipf exponent is estimated using maximum likelihood (as suggested by [52]; cf. Materials & Methods for further details) both for the complete data and for a truncated version of the data where word types that occur less than 20 times are excluded for the BNC data and for the DeReKo data (I chose this cut-off point instead of 40 since the GNgC data is of course much bigger than the corpora used in this section). It is important to note that in both cases the Zipf parameter is calculated using a random sample of approximately 1,000,000 word tokens of the data, because as argued above, the Zipf parameter depends on the size of the corpus [27]. To account for random noise, the analyses are repeated 20 times. Figure **4** shows that there is an obvious difference between the Zipf parameter for the untruncated and the truncated version of the data, however this difference does not seem to vary with the size of the corpus but is relatively constant, excluding what appears to be random noise. Separate analyses of variance for the BNC and DeReKo data confirm

that in both cases only the version type (truncated vs. untruncated) is an important factor (*p*s < .0001) to predict the value of the Zipf exponent, and not the sample size or the interaction of both variables (all *p*s > .1). The effect of the data truncation is opposed for both corpora: while truncation decreases the value of the exponent for the BNC data, it increases the value of the exponent for the DeReKo data. The most likely explanation for this result is the fact that the BNC corpus is much smaller than the DeReKo corpus.

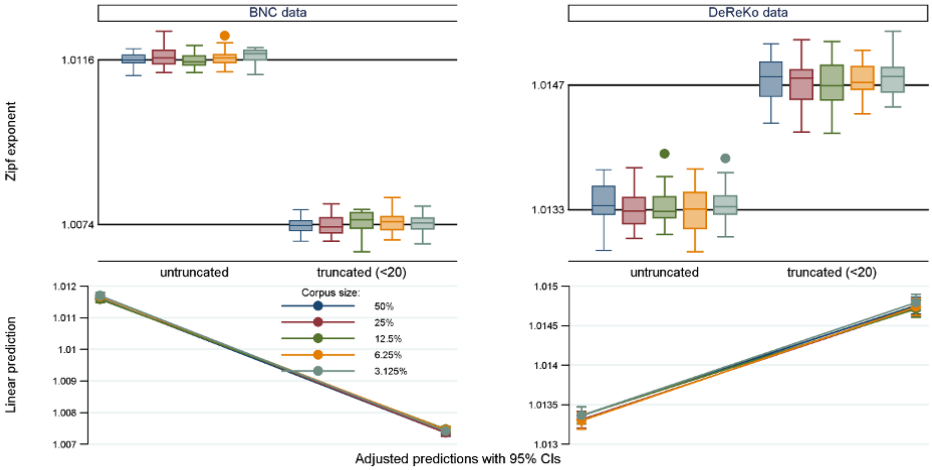

Figure 4: The relationship of the corpus size and the Zipf exponent with or without data truncation. Both the BNC and the DeReKo data are divided into independent sub-corpora varying in size (50%, 25%, 12.5%, 6.25%, 3.125%) and for each sub-corpus the Zipf exponent is estimated using maximum likelihood. The plots on the top-side show that the size of the Zipf exponent varies significantly ($p < .0001$) if the data is truncated (mean values on the y-axes). In this context, neither sample size nor interaction of sample size and version type (truncated vs. untruncated) are significant predictors of the size of the Zipf exponent (*p*s > .1) as the margins plots on the bottom-side demonstrate the results of a two-way ANOVA of the size of the Zipf exponent on the version type, sample size, and their interaction. This indicates that is is possbile to deduce overall trends on the basis of GNgC data, because the data truncation in interaction with the corpus size does not systematically bias the results.

Nevertheless, the results presented above imply that it is possible to draw conclusions about general trends, since the effect of data truncation does not seem to systematically interact with the size of the corpus. To further investigate this assumption, I calculated the Zipf exponent on the actual GNgC data both based on the 1-gram and on the aggregated 2-gram data. The rationale for this procedure is very simple: if the results based on the 2-gram data are systematically different from the results

based on the 1-gram data, one cannot rule out the possibility that the true data, that is the data without truncation, is also systematically different from the 1-gram data.

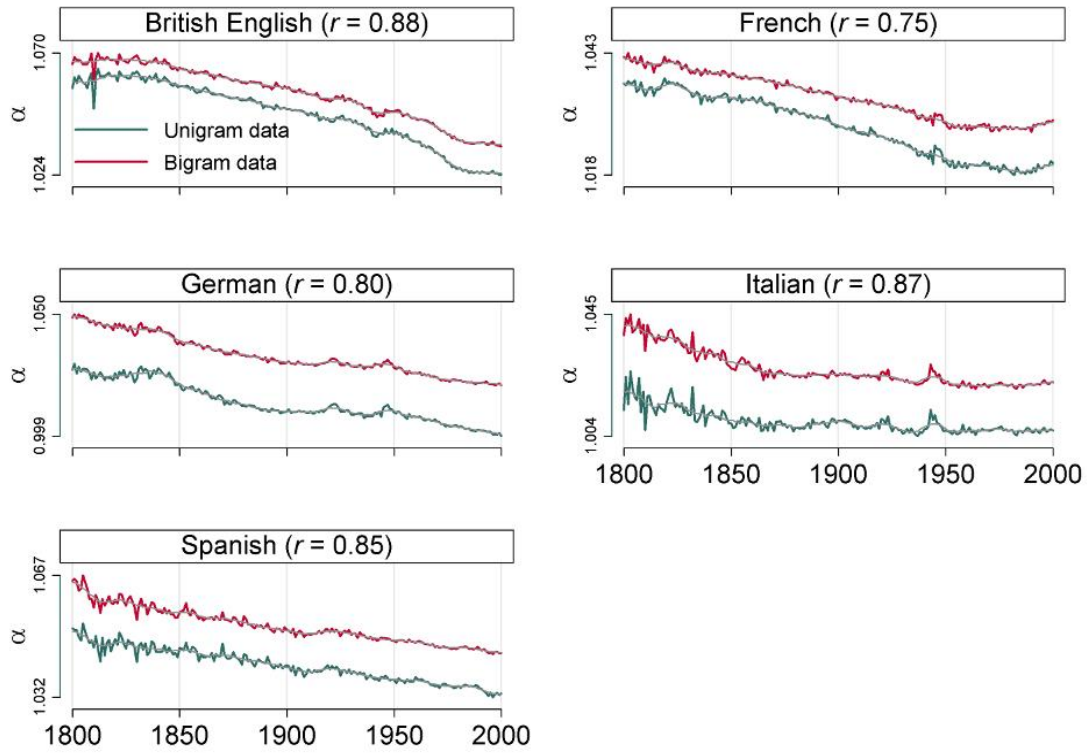

Figure 5: The size of the Zipf exponent (α) as a function of time for five different languages. Emerald lines: maximum likelihood estimation of the Zipf exponent based on the token frequencies of the 1-gram GNgC data. Cranberry lines: analogous maximum likelihood estimation of the Zipf exponent based on the 2-gram data. The correlation values are the Lag 0 correlations for the first differences (accounting for non-stationarity) of the values for the 1-gram and the 2-gram data. The strong correlation indicates that data truncation does not systematically bias the results due to the changing corpus size as a function of time. The change of the Zipf exponent for all investigated languages points towards a change in linguistic complexity as it is represented in the GNgC.

However, if the resulting time series of the 1-gram and the 2-gram data are similar, then one can at least cautiously deduce an overall trend. Figure **5** shows the resulting plots for all investigated languages. Since it is the main purpose of this analysis to show how to handle the GNgC data truncation, I will only briefly interpret the actual results: in all five languages, the Zipf exponents (both based on the 1-gram and on the 2-gram data) show an overall tendency to decrease as a function of time. This change is not mainly driven by the changing corpus size, as the weak correlations between the first differences of the Zipf exponent and the corpus size reveal (r_{GB} = 0.05;

r_{FRE} = 0.22; r_{GER} = 0.05; r_{ITA} = 0.13; r_{SPA} = 0.11). Hence, the change of the Zipf exponent seems to point towards a change in linguistic complexity. In a future publication, I will investigate how this relates to other measures of lexical and syntactical complexity.

Regarding the problem of data truncation, the different calculations based on the 1-gram and the 2-gram data point in the same direction. The correlation values are the Lag 0 correlations for the first differences (accounting for non-stationarity) of the values for the 1-gram and the 2-gram data. The strong correlation indicates that data truncation does not systematically bias the results due to the changing corpus size as a function of time. Therefore, it seems appropriate to deduce a negative overall trend for all investigated languages, which points towards a change in linguistic complexity as it is represented in the GNgC data.

Conclusion

To avoid breaking any copyright laws, the Culturomics team opted for a twofold strategy: a) the GNgC datasets are not accompanied by any metadata regarding the texts the corpora consist of, b) the data are truncated to prevent an indirect conclusion from the n-gram to the author of the text. The consequences of this strategy were analyzed in this paper.

One could ask the question whether the problems raised in this paper imply that the GNgC data cannot be used to measure cultural and linguistic change *at all*. This is certainly not the case. A freely available database which is that big and comprehensive just does not exist anywhere else and it seems obvious that only a project with as big a company as Google in the background would be able to accomplish this truly Herculean task. It goes without saying that legal restrictions also have to be taken seriously in this case. Maybe different strategies will be devised to overcome the problems presented in this paper for future versions of the datasets. For example, the problem of data truncation could be worked around, without breaking any legal restrictions by using e.g. a pseudo code for the n-grams that occur less than 40 times. This in turn would make it absolutely possible to

use the datasets for macroscopic analyses like the evolution of the Zipf exponent presented above, where actual words only play a subordinate role.

Generally, the impact of lacking metadata and data truncation is hard to estimate. It most certainly varies from case to case and it is unquestioned that a significant amount of the cultural and linguistic reality is encoded in the data. Nevertheless, I believe that the high standards that are prevalent in the empirical sciences imply that due to the problems outlined above, conclusions based on the GNgC have to be treated very cautiously. As long as the problems persist, I believe that it is too early to assess to what extent the GNgC data really enable researcher to extend "the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities" as [2, p. 1] put it. Instead of speaking about general linguistic or cultural change, it seems preferable to explicitly restrict any results to linguistic or cultural change *as it is represented in the Google Ngram data*. For all kinds of research with language data - to rephrase a quote by [31, p. 249] - it is important to realize that size cannot make up for a lack of metadata.

Material and Methods

All analyses were carried out using Stata/MP2 12.1 for Windows (64-bit version).

I used the full datasets that were made available by [2] at [1]. For the present study, both the 1-gram and the 2-gram datasets of Version 2 (July 2012) of the following languages were used: British English, French, German, Italian and Spanish. To make the results presented in this study reproducible, S1 contains Stata programs (“do files”) to read the unzipped raw data and prepare cleaned corpora for each year (in case of the 2-gram data) and one big dataset (in case of the 1-gram data). All 1-gram corpora share the same basic structure, in which the first column is the string variable for the word, the second variable contains the word-class (POS) information as described in [3] and the third column contains the match count for one particular year (e.g. match1899). For the 2-gram corpora, the structure is similar and contains a string variable and word class information both for the first and for the second word of one 2-gram. The datasets are very large, so this step

took several weeks. For example, reading in the British English data on a multicore 2.00GHz processor with 82GB available RAM took more than three weeks to finish.

The Nazi censorship examples are based on the German 2-gram data. A Stata program to reproduce the examples and plots can be found in S2. 2-grams in which the first word indicates the beginning of a sentence and 2-grams in which the second word indicates the end of a sentence [34] were excluded from the analysis. To compute the relative frequencies, the number of matches in a year were divided by the total number of matches in that year and multiplied by 1,000,000. Thus the match-counts represent the number of occurrences of one particular name per one million word tokens. The time-series for the individuals were smoothed using a simple weighted moving average with an eleven-year window centered on the current match-count [57].

As mentioned in the text, the calculation of average book-lengths used the total-counts files for each investigated language. Again the time-series were smoothed with an eleven-year window. The Stata program to reproduce Figure **2** can be found in S3 of the Supporting Information.

A Stata program that calculates the 1-gram token frequencies using the 2-gram data sets and reproduces Table **1** can be found in Appendix S4. Very surprisingly, word types for the British English and the German datasets that only occur in the 2-gram data but not in the 1-gram data also seem to exist. According to [36] this cannot be the case. However, a visual inspection of the raw GNgC data reveals that this indeed seems to be the case.

Figure **4** in which the time series of the relative frequency of *hapax legomena* is cross-correlated with the time series of the corpus size can be reproduced using the Stata program in Appendix S5. A Phillips-Perron [58] test was carried out both for the time series of the relative frequency of *hapax legomena* and for the time series of the corpus size for each language. In each case, a unit root was confirmed, indicating non-stationarity (all MacKinnon approximate *p*s > .01). Taking the first differences of the series results in (weakly) stationary series in all cases (all MacKinnon approximate *p*s < .01).

The BNC frequency list was compiled by [54] and is available for download here [59]. The DeReKo list is copyright protected and cannot be published. Further information, for example on available metadata can be found in [55]. Both the BNC and the DeReKo lists were modified according to the special GNgC design rules used by [36, II.3.]. For example, tokens that contain a hyphen are treated as three separate words. As mentioned in the text, the effect of the data truncation of the size of the Zipf exponent is opposed for both kinds of data. In the article, this was attributed to the fact that that the BNC corpus is much smaller than the DeReKo corpus. However, this could also possibly imply that the data in the DeReKo data that is excluded due to the truncation process, represents a different kind of information compared to the BNC data. For example, complex compound words in German are created by combining individual word strings to one word string (e.g. “Korpusgröße”), while similar string chains in English usually feature hyphens or spaces (e.g. “corpus size”). Therefore truncating words that occur less than 10 times eliminates almost 77% of all word types for the BNC data, but – on average – almost 87% for the DeReKo samples. Further analyses (not reported here) reveal that the difference has nothing do to with the fact that the BNC frequency list is not case sensitive. This is an avenue for future research.

Mathematically, Zipf’s law can be modeled as a right-truncated zeta distribution [52], where the probability *p* of a word with rank *r* is:

$$p(r) = \frac{1}{\sum_{r=1}^{N} r^{-\alpha}} r^{-\alpha} \qquad (4)$$

Here, *N* is the observed number of word types, that is, the vocabulary size in a given sample. A Stata program that implements the GNgC design rules, fits the α exponent by a maximum likelihood estimation procedure as described by [52]. Reproductions of Figure **4** and the ANOVA can be found in Appendix S6.

For the maximum likelihood estimation of the Zipf exponent for each year and each investigated language, the initial value of α at t_0 was set to the estimated value of α at t_{-1}. The maximum number

of ML iterations to converge was set at 1000. A Stata program can be found in S7 of the Supporting Information. It draws random corpus samples approximately sized one million word tokens, ML estimates the Zipf exponent and reproduces Figure **5** . It is worth pointing out that the trends of the individual languages may seem to correlate, but are not related in any substantial way, as the weak inter-languages correlations of the first differences of the Zipf exponent indicate ($|r| < 0.17$ for all comparisons except the correlation between British English and Italian ($r = 0.38$); further analyses reveal that the period from 1800-1814 significantly influenced this result. If the time span is restricted to the period of 1815-2000 then all $|r|$ are below 0.17).

Supporting Information

Text SI1 contains all Stata programs ("do-files") that are mentioned in the Material & Methods section that can be used to reproduce the figures and calculations presented in the text (starting with S1 and then numbered consecutively).

Text SI2 contains time series plots similar to those presented in Figure 1 for both the 50 most suppressed people and the 50 most enhanced persons according to [39]**.**

Acknowledgments

I wish to thank my colleagues at IDS, Stefan Engelberg, Marc Kupietz, Carolin Müller-Spitzer and Sascha Wolfer for all the helpful discussions regarding the topics presented in this paper. I thank Heike Stadler for providing me with the DeReKo frequency list and Sarah Signer for proofreading the draft version of this paper. I also wish to thank the IDS for generously providing me with all the hard- and software equipment I needed in order to analyze the data. Finally I am grateful to my colleague Peter Meyer for the countless discussions on general linguistic topics and for mathematical and quantitative advice. Unfortunately, all remaining errors are mine.

References

1. www.culturomics.org (2014). Available: http://www.culturomics.org/. Accessed 10 March 2014.

2. Michel J-B, Shen YK, Aiden AP, Verses A, Gray MK, et al. (2010) Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331: 176–182 [online pre–print: 1–12]. doi:10.1126/science.1199644.

3. Lin Y, Michel J-B, Aiden LE, Orwant J, Brockmann W, et al. (2012) Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea. pp. 169–174.

4. Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The Expression of Emotions in 20th Century Books. PLoS ONE 8: e59030. doi:10.1371/journal.pone.0059030.

5. Bentley RA, Acerbi A, Ormerod P, Lampos V (2014) Books Average Previous Decade of Economic Misery. PLoS ONE 9: e83147. doi:10.1371/journal.pone.0083147.

6. Gao J, Hu J, Mao X, Perc M (2012) Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. J R Soc Interface 9: 1956–1964. doi:10.1098/rsif.2011.0846.

7. Gulordava K, Baroni M (2011) A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. Edinburgh, Scotland: Association for Computational Linguistics. pp. 67–71.

8. Juola P (2013) Using the Google N-Gram corpus to measure cultural complexity. Lit Linguist Comput 28: 668–675. doi:10.1093/llc/fqt017.

9. Kesebir P, Kesebir S (2012) The cultural salience of moral character and virtue declined in twentieth century America. J Posit Psychol 7: 471–480. doi:10.1080/17439760.2012.715182.

10. Lieberman E, Michel J-B, Jackson J, Tang T, Nowak MA (2007) Quantifying the evolutionary dynamics of language. Nature 449: 713–716. doi:10.1038/nature06137.

11. Perc M (2012) Evolution of the most common English words and phrases over the centuries. J R Soc Interface 9: 3323–3328. doi:10.1098/rsif.2012.0491.

12. Petersen AM, Tenenbaum J, Havlin S, Stanley HE (2012) Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. Sci Rep 2. Available: http://www.nature.com/doifinder/10.1038/srep00313. Accessed 10 March 2014.

13. Petersen AM, Tenenbaum JN, Havlin S, Stanley HE, Perc M (2012) Languages cool as they expand: Allometric scaling and the decreasing need for new words. Sci Rep 2. Available: http://www.nature.com/doifinder/10.1038/srep00943. Accessed 10 March 2014.

14. Ravallion M (2011) The Two Poverty Enlightenments: Historical Insights from Digitized Books Spanning Three Centuries. Poverty Public Policy 3: 167–212. doi:10.2202/1944-2858.1173.

15. Twenge JM, Campbell WK, Gentile B (2012) Increases in Individualistic Words and Phrases in American Books, 1960–2008. PLoS ONE 7: e40181. doi:10.1371/journal.pone.0040181.

16. Aiden E (2013) Uncharted: big data as a lens on human culture. New York: Riverhead Hardcover, A member of Penguin Group (USA).

17. Bochkarev V, Solovyev V, Wichmann S (2014) Universals versus historical contingencies in lexical evolution. Available: http://wwwstaff.eva.mpg.de/%7Ewichmann/LexEvolUploaded.pdf. Accessed 12 June 2014.

18. Gerlach M, Altmann EG (2013) Stochastic Model for the Vocabulary Growth in Natural Languages. Phys Rev X 3. Available: http://link.aps.org/doi/10.1103/PhysRevX.3.021006. Accessed 12 June 2014.

19. Frühwald J (2012) Don't worry, I'm a physicist. Val Syst. Available: http://val-systems.blogspot.de/2012/07/dont-worry-im-physicist.html. Accessed 10 March 2014.

20. Jockers ML (2010) Unigrams, and bigrams, and trigrams, oh my. Available: http://www.matthewjockers.net/2010/12/22/unigrams-and-bigrams-and-trigrams-oh-my/. Accessed 10 March 2014.

21. Jockers ML (2013) Macroanalysis: digital methods and literary history. Urbana: University of Illinois Press. 192 p.

22. Liberman M (2012) Textual narcissism. Lang Log. Available: http://languagelog.ldc.upenn.edu/nll/?p=4069. Accessed 10 March 2014.

23. Liberman M (2012) Textual narcissism, replication 2. Lang Log. Available: http://languagelog.ldc.upenn.edu/nll/?p=4071. Accessed 10 March 2014.

24. Liberman M (2013) Word String frequency distributions. Lang Log. Available: http://languagelog.ldc.upenn.edu/nll/?p=4456. Accessed 10 March 2014.

25. Schmidt B (2013) Are words the atomic unit of a dynamic system? Sapping Atten. Available: http://sappingattention.blogspot.de/2013/02/are-words-atomic-unit-of-dynmic-system.html. Accessed 10 March 2014.

26. Underwood T (2012) ngrams | The Stone and the Shell. Stone Shell - Hist Quest Rais Quant Approach Lang. Available: http://tedunderwood.com/category/ngrams/. Accessed 10 March 2014.

27. Baayen RH (2001) Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers.

28. Hans-Jörg Schmid (2010) Does frequency in text instantiate entrenchment in the cognitive system? In: Dylan Glynn, Kerstin Fischer, editors. Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches. Berlin, New York: de Gruyter. pp. 101–133.

29. Baroni M, Evert S (2009) Statistical methods for corpus exploitation. In: Lüdeling A, Kytö M, editors. Corpus linguistics: An international handbook. Berlin: De Gruyter Mouton, Vol. 2. pp. 777–802.

30. Johansson S (2009) Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: Lüdeling A, Kytö M, editors. Corpus linguistics: an international handbook. Handbücher zur Sprach- und Kommunikationswissenschaft = Handbooks of linguistics and communication science. Berlin ; New York: Walter de Gruyter. pp. 33–55.

31. Biber D (1998) Corpus linguistics: investigating language structure and use. Cambridge ; New York: Cambridge University Press. 300 p.

32. Biber D (1991) Variation across speech and writing. Cambridge [England]; New York: Cambridge University Press.

33. Lee DYW (n.d.) Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. Lang Learn Technol 5: 37–72.

34. Biber D (2012) Register as a predictor of linguistic variation. Corpus Linguist Linguist Theory 8: 9–37.

35. Biber D, Gray B (2013) Being Specific about Historical Change : The Influence of Sub-Register. J Engl Linguist. Available: http://eng.sagepub.com/cgi/doi/10.1177/0075424212472509. Accessed 14 April 2014.

36. Michel J-B, Shen YK, Aiden AP, Verses A, Gray MK, et al. (2010) Quantitative Analysis of Culture Using Millions of Digitized Books (Supporting Online Material). Science 331. Available: http://www.sciencemag.org/content/early/2010/12/15/science.1199644/suppl/DC1. Accessed 5 March 2014.

37. documentArchiv.de - Wehrgesetz (21.05.1935) (n.d.). Available: http://www.documentarchiv.de/ns/1935/wehrgesetz.html. Accessed 15 May 2014.

38. Jurafsky D, Martin JH (2009) Speech and Language processing: an introduction to natural language processing, computational Linguistics, and speech recognition. Upper Saddle River: Pearson Education (US).

39. Supporting Online Material (n.d.). Available: http://www.sciencemag.org/content/331/6014/176/suppl/DC1. Accessed 15 May 2014.

40. Biographie: Theodor Blank , 1905-1972 (n.d.). Stift Haus Gesch Bundesrepub Dtschl. Available: http://www.hdg.de/lemo/html/biografien/BlankTheodor/index.html. Accessed 15 May 2014.

41. Zinn H (2001) A people's history of the United States: 1492-present. New York: Perennial Classics.

42. Bertin C, Petersdorff C von (1989) Die letzte Bonaparte: Freuds Prinzessin : ein Leben. Freiburg i. Br: Kore.

43. Frederiksen BF (2008) Jomo Kenyatta, Marie Bonaparte and Bronislaw Malinowski on Clitoridectomy and Female Sexuality. Hist Workshop J 65: 23–48. doi:10.1093/hwj/dbn013.

44. Huebener RP, Lübbig H (2012) A focus of discoveries. 2nd ed. Singapore ; Hackensack, N.J: World Scientific. 185 p.

45. Tweedie FJ, Baayen RH (1998) How Variable May a Constant be? Measures of Lexical Richness in Perspective. Comput Humanit 32: 323–352.

46. Baroni M (2009) Distributions in text. In: Lüdeling A, Kytö M, editors. Corpus linguistics: an international handbook. Handbücher zur Sprach- und Kommunikationswissenschaft = Handbooks of linguistics and communication science. Berlin ; New York: Walter de Gruyter. pp. 803–821.

47. Zipf GK (2002) The psycho-biology of language ; an introduction to dynamic philology. Available:

http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=660639. Accessed 10 March 2014.

48. Zipf GK (2012) Human behavior and the principle of least effort: an introduction to human ecology. Mansfield Centre, CT: Martino Pub.

49. Ha LQ, Sicilia-Garcia EI, Ming J, Smith FJ (2002) Extension of Zipf's law to words and phrases Association for Computational Linguistics, Vol. 1. pp. 1–6. Available: http://portal.acm.org/citation.cfm?doid=1072228.1072345. Accessed 2 May 2014.

50. Kilgarriff A (2005) Language is never, ever, ever, random. Corpus Linguist Linguist Theory 1. Available: http://www.degruyter.com/view/j/cllt.2005.1.issue-2/cllt.2005.1.2.263/cllt.2005.1.2.263.xml. Accessed 2 May 2014.

51. Bybee J (2003) Mechanisms of Change in Grammaticization: The Role of Frequency. In: Joseph BD, Janda RD, editors. The Handbook of Historical Linguistics. Oxford, UK: Blackwell Publishing Ltd. pp. 602–623. Available: http://doi.wiley.com/10.1002/9780470756393.ch19. Accessed 2 May 2014.

52. Baixeries J, Elvevåg B, Ferrer-i-Cancho R (2013) The Evolution of the Exponent of Zipf's Law in Language Ontogeny. PLoS ONE 8: e53227. doi:10.1371/journal.pone.0053227.

53. Bentz C, Kiela D, Hill F, Buttery P (2014) Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. Corpus Linguist Linguist Theory 0. Available: http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2014-0009/cllt-2014-0009.xml. Accessed 12 May 2014.

54. Kilgarriff A (1997) Putting frequencies in the dictionary. Int J Lexicogr 10: 135–155.

55. Kupietz M, Belica C, Keibel H, Witt A (2010) The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari N, Tapias D, Rosner M, Piperidis S, Odjik J, et al., editors. Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10). Valetta, Malta: European Language Resources Association (ELRA). pp. 1848–1854.

56. Piantadosi ST (2014) Zipf's word frequency law in natural language: A critical review and future directions. Psychon Bull Rev. Available: http://link.springer.com/10.3758/s13423-014-0585-6. Accessed 2 May 2014.

57. Becketti S (2013) Introduction to time series using Stata. 1st ed. College Station, Tex: Stata Press. 443 p.

58. Phillips PCB, Perron P (1988) Testing for a unit root in time series regression. Biometrika 75: 335–346. doi:10.1093/biomet/75.2.335.

59. Read-me for Kilgarriff's BNC word frequency lists (n.d.). Available: http://www.kilgarriff.co.uk/bnc-readme.html. Accessed 15 May 2014.