Alexander Koplenig

Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis

Abstract: Using the Google Ngram Corpora for six different languages (including two varieties of English), a large-scale time series analysis is conducted. It is demonstrated that diachronic changes of the parameters of the Zipf–Mandelbrot law (and the parameter of the Zipf law, all estimated by maximum likelihood) can be used to quantify and visualize important aspects of linguistic change (as represented in the Google Ngram Corpora). The analysis also reveals that there are important cross-linguistic differences. It is argued that the Zipf–Mandelbrot parameters can be used as a first indicator of diachronic linguistic change, but more thorough analyses should make use of the full spectrum of different lexical, syntactical and stylometric measures to fully understand the factors that actually drive those changes.

Keywords: Zipf's law, Zipf–Mandelbrot law, power law, lexical richness, vocabulary size, type token ratio, syntactic complexity, noun–pronoun ratio, diachronic corpus linguistics, time series analysis, Google Ngram Corpora

DOI 10.1515/cllt-2014-0049

1 Introduction

Phenomena where f is a quantity that is drawn from the following probability distribution:

$$p(f) \propto f^{-\alpha} \qquad [1]$$

are said to be power-law distributed. Many different empirical phenomena have been proposed to be characterizable by this particular kind of distribution: the population of cities, the intensity of wars, the number of species per genus of

Alexander Koplenig, Institute for the German language (IDS), Mannheim, Germany,
E-mail: koplenig@ids-mannheim.de

mammals or the numbers of copies of bestselling books (cf. Newman 2005; Clauset et al. 2009 for two seminal papers in this area)[1]. The properties of this distribution have some interesting consequences, most notably – compared for example to the famous normal distribution – its heavy tail. Clauset et al. (2007) provide a nice example:

> [C]onsider a world where the heights of Americans were distributed as a power law, with approximately the same average as the true distribution (which is convincingly normal when certain exogenous factors are controlled). In this case, we would expect nearly 60 000 individuals to be as tall as the tallest adult male on record, at 2.72 meters. Further, we would expect ridiculous facts such as 10 000 individuals being as tall as an adult male giraffe, one individual as tall as the Empire State Building (381 meters), and 180 million diminutive individuals standing a mere 17 cm tall (Clauset et al. 2007: 6).

Maybe the most cited distributions that are thought to follow a power law are word frequency lists, which tend to be Zipf-distributed (Zipf 1935, Zipf 2012). If one assigns rank 1 to the most frequent word (type), 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:

$$f(r) \propto r^{-\alpha} \qquad [2]$$

The exponent α is a parameter that has to be determined empirically. In the simplest case, α is equal to unity (as Zipf frequently assumed). Figure 1 presents the standard way to plot a word frequency distribution where both the horizontal and the vertical axis are logarithmic. For the left side of Figure 1, an unlemmatized word frequency list for British English compiled by Kilgarriff (1997) of the written part of the British National Corpus (BNC), a roughly 90-million word collection was used. For the right side, an unlemmatized word frequency list for German taken from the German Reference Corpus DeReKo (Version 2011) that consists of roughly 4.5 billion word tokens was used (for details on corpus design and compilation see Kupietz et al. 2010).

This property of word frequency distributions can be used to understand different linguistic phenomena. In a recent study for example, Yang (2013) uses the statistical properties of Zipf's law to calculate a statistical profile of grammar and shows how this can accurately explain the low-syntactic diversity observed in language use.

In the present article, the statistical properties of this well-known fact about languages are used to measure a second – equally well-known – fact about

[1] Newman (2005) argues that the actual mechanisms that lead to power-law distributions are likely to be different for different phenomena.

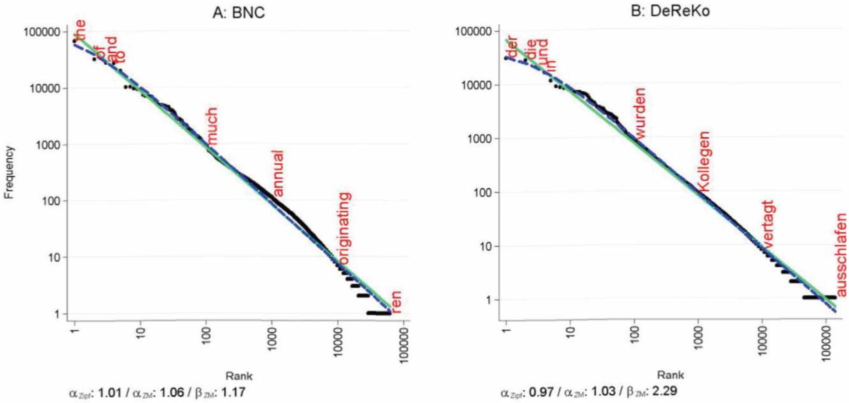


Figure 1: Two word frequency distributions (1 million samples in each case). Plot A: data from the British National Corpus. Plot B: data from the German Reference Corpus. The black dots represent the empirical data; the solid mint line represents the fit to a Zipf distribution; the blue dashed line represents the fit to a Zipf–Mandelbrot distribution. The fitting and sampling procedure is described in the text. The parameters of the distributions are added as a note. Some examples are depicted in red.

languages, namely that languages are constantly changing on all fundamental levels (Labov 1994). For example, Baixeries et al. (2013) fit longitudinal data from the CHILDES database (MacWhinney 2014) to a right-truncated zeta distribution and estimate the exponent of the Zipf law by maximum likelihood. The analysis reveals that the exponent of the law clearly tends to decrease over time for children (and less clearly for adults) which points toward the fact that the exponent of the law is an indicator of linguistic complexity.

Mandelbrot (1953) modified the original Zipf law by adding a second free parameter β:

$$f(r) \propto (r + \beta)^{-\alpha} \qquad [3]$$

where the original Zipf law is a special case with β set to 0. This modification accounts for the fact that when the log of the frequency of word types is plotted against the log of its rank, word types with low ranks tend to deviate from the observed linearity between frequency and rank (cf. the blue dashed lines in Figure 1). Therefore, changes of the β parameter approximate changes of the class of function words, since those words, for example pronouns, determiners, prepositions or conjunctions, tend to have low ranks. Correspondingly, Bentz et al. (2014a) use the parameters of the Zipf–Mandelbrot law (ZM) to classify languages according to what they call a “grammatical fingerprint”. According to the authors, this fingerprint can help to understand changing morphological encoding strategies. For example, they show that for the English language, changes of the ZM parameters approximate the loss of morphological marking as an important factor for changing word frequency distributions. In a second study, Bentz et al. (2014b) show that for cross-linguistic comparisons, higher values of the ZM parameters are associated with lower lexical diversity.

The aim of this paper is to put the information encoded by the parameters of ZM law further into perspective. To this end, I use the large-scale diachronic Google Ngram Corpora (GNgC, Michel et al. 2010a) for six different languages (including two varieties of English). For each investigated language and for each year between 1800 and 2000, the ZM α parameter is estimated by maximum likelihood[2]. In addition, three further measures of linguistic variation are calculated: the vocabulary size as the most fundamental measure of lexical richness (Tweedie and Baayen 1998), the mean sentence length as a measure of syntactic complexity (Szmrecsanyi 2004) and the noun–pronoun ratio as a measure indicative of stylistic tendencies (Säily et al. 2011). The resulting different time series are then correlated with the α_{ZM} parameter for each investigated language.

The remainder of this contribution is structured as follows: First, the data sets and the creation of diachronic corpora will be presented (Section 2). In the next section, some important methodological considerations for the analysis of time series data are briefly described (Section 3). This section is followed by the operationalization of the indicators mentioned above (Section 4). The methods to analyze the data are described in Section 5. In Section 6, the results are presented and discussed. This paper ends with some concluding remarks in Section 7.

2 The data

In this paper, the full datasets, made available by Michel et al. (2010a) at www.culturomics.org, were used[3]. For the present study, both the 1-gram and the 2-gram datasets of Version 2 (July 2012) of the following languages were used: American English, British English, French, German, Italian and Spanish. 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

[2] For reasons of clarity, this analysis focusses on the shape parameter of the ZM law. In Appendix A.3, additional results are presented where the analysis is re-run for both the β-parameter of the ZM law, for the α-parameter of the Zipf law and for power laws.

[3] Last accessed on 8 September 2014.

described in Lin et al. (2012) 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.[4]

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. Therefore, I believe that conclusions based on the GNgC have to be treated very cautiously. In particular, this means that without proper metadata, it is hardly possible to separate (general) language internal and language external (conventional or environmental) change (Szmrecsanyi 2014), or changes that reflect register differences in the composition of a corpus (Biber 1991; Biber and Gray 2013). On these grounds, all results presented in this paper are explicitly restricted to linguistic (and cultural) change *as it is represented in the Google Ngram data.*

It is equally important to say a few words about the problem of text-length dependence, which is quite well known in corpus linguistics. Among others, Tweedie and Baayen (1998) demonstrate very convincingly that any of the quantitative measures proposed to study lexical richness varies with increased corpus size. Given the fact that the corpus size based on the GNgC for each year (and for each language) strongly increases as a function of time, calculating the variables based on the actual corpus sizes would systematically bias the results. To solve this problem, an efficient and computationally cheap method is to draw random samples of 1,000,000 tokens from the data by performing a binomial split for each year (as suggested by Piantadosi 2014). For each word type w, this procedure returns binomial (n_{wy},p_y) random variates, where n_{wy} is the raw token frequency of the word type w in the year y and p is the success probability in year y, which is given as: $(1{,}000{,}000 + 10{,}000)/s_\mathrm{y}$, where s_y is the corpus size in the year y.[5] The resulting corpora sized 1,000,000 token are what Tweedie and Baayen (1998) would call a full randomized sample of all texts in a given year. Of course, quantitative investigations regarding the actual discourse structure are not possible with randomized samples of this kind, but would not be

[4] The datasets are very large, so this step took several weeks. For example, reading the British English data on a multicore 2.00 GHz processor with 82 GB available RAM took more than three weeks to finish. All analyses were carried out using Stata/MP2 12.1 for Windows (64-bit version). All Stata do-files are available upon request.

[5] Since this process is random *per definition*, instead of using 1,000,000 as the nominator for s_y, 1,010,000 million was used to obtain a sample which is slightly bigger than 1,000,000 tokens. To generate a sample of exactly 1,000,000 tokens, all drawn tokens where thrown in an urn from which 1,000,000 million tokens were then drawn randomly.

possible anyway on the basis of the GNgC that only contain token frequencies for each n-gram type.[6]

3 Analyzing time series data

The statistical analysis of time series data, that is data with a natural temporal ordering, is special. In fact it is so special that most of the classical statistical tools of data analysis cannot be used directly. There are two (or three[7]) main reasons for this. The first reason is the sequential dependence of observations. This fundamental property of time series means that the errors of observations that are close together in time tend to be (auto-)correlated. This in turn violates one of the assumptions of OLS (ordinary least square) regressions and produces estimators that are not BLUE (Best Linear Unbiased Estimate). The second reason is that many time series have a unit root, which means that the time series variable is non-stationary. Or put differently, the variable which is measured at successive moments in time exhibits an upward or downward trend. Regressing one non-stationary time series on another non-stationary time series leads to a spurious model where the variables look highly correlated but are not related in any substantial sense (Granger and Newbold 1974). To demonstrate this, I simulated two random walks with drift (Hill 2008; Becketti 2013: 387/389), where the value x_t (and y_t) at time point t is given as:

$$x_t = 0.09 + x_{t-1} + e_t \tag{4}$$

with e_t normally distributed in the interval [0,1]. This means that the resulting time series x and y both have an average upward trend but are statistically

[6] For the sampling procedure, tokens tagged as numerals and punctuation were excluded. To account for some obvious errors in the tokenization and tagging of POS (Michel et al. 2010b), the following manipulations were performed: <d'> and <D'> were POS tagged as adpositions (except when the word was POS tagged as a determiner) and the apostrophe was removed. < l' > and < L' > were tagged as determiners (except when the word was tagged as a pronoun) and the apostrophe was removed. For the Spanish data, <á> was tagged as an adposition and <ó> was tagged as a conjunction. Words that were longer than five characters and did not contain at least one alphanumeric character (regular expression: [A-Za-z0-9]) were excluded (e.g. ******, ... , ————, _______). Strings consisting solely of the following characters were removed, too: « » . ' * §• .. ° # $. + ^* ( ) [ ] {} - = | \ : ; < , > ? / ~ `". Finally, words consisting of only numeric characters were excluded.

[7] The third reason has to with the fact that when analyzing cross-sectional data, we assume that we can estimate population means using the sample means. In time series analysis, however, the population mean does not necessarily exist at all (Chatfield 2004: Ch. 4).

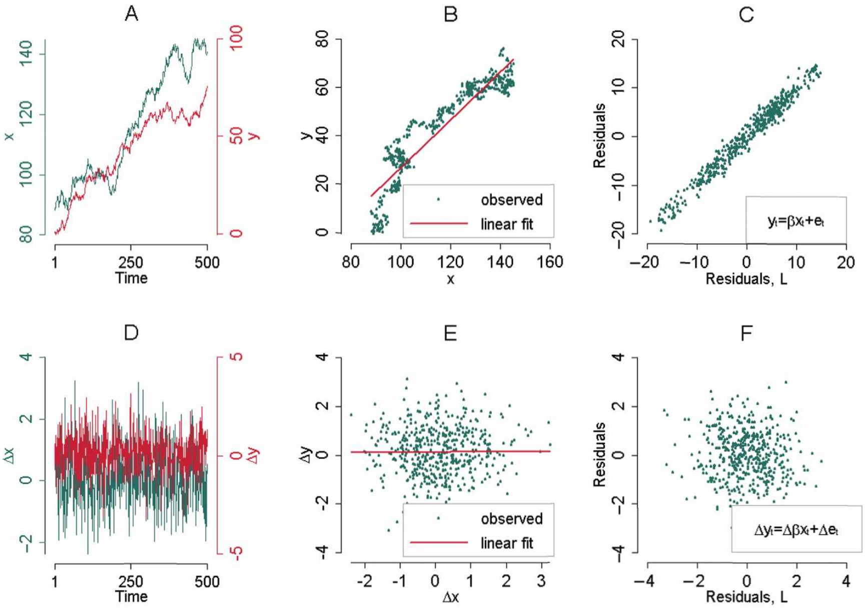


Figure 2: Two random walks (*x* and *y*) with drift (plot A). A Scatterplot (B) of the two random walks suggests that *x* and *y* are strongly related. Therefore, the true hypothesis that there is no correlation between *y* and *x* would be rejected. Plot C shows that the residuals are strongly auto-correlated. To solve this problem, plot D compares the first differences of *x* and *y*. Graph E demonstrates that there is no significant linear relationship between Δ*x* and Δ*y*, while plot F shows no sign of autocorrelation of the residuals.

completely unrelated. Plot A of Figure 2 graphs the two series, while graph B plots *y* against *x*. The line in plot B plots the result of an OLS regression of *y* on *x*, using the following standard notation:

$$y_t = \beta x_t + e_t \qquad [5]$$

In this case, the *t*-value is 54.46 and highly significant ($p < 0.001$). Apparently, the model fit is also very good, as indicated by an R^2 of 0.86. Therefore the true hypothesis that there is no correlation between *y* and *x* would be rejected (Becketti 2013: 387–389). However, we find that the Durbin–Watson statistic is very low (0.03), which points toward the fact that the residuals of the regression analysis are heavily auto-correlated, because in the presence of no autocorrelation it should be somewhere around 2 (Granger and Newbold 1974; Hamilton 2013: 370–371). Plot C of Figure 2 plots the residuals against the lagged residuals demonstrating a strong autocorrelation of the residuals which means that the model is severely mis-specified as result of non-stationarity.

There are two main recommendations to induce stationarity (Granger and Newbold 1974). The first one is to include a lagged dependent variable as an independent variable in the regression eq. [6], that is, fit a dynamic model as:

$$y_t = \beta_1 x_t + \beta_2 y_{t-1} + e_t \tag{6}$$

If we regress y on x and include a lagged y in the example mentioned above then the t-value of β_1 becomes small (0.70) and does not reach statistical significance ($p = 0.49$). In addition to that, the Durbin–Watson statistic is much higher (2.16) compared to the version with no lagged term, indicating a better model specification.

To obtain (weakly) stationary series in this paper, I follow the second general recommendation: instead of comparing the actual variables, I take first differences of the variables involved which are defined as:

$$\Delta x_t = x_t - x_{t-1} \tag{7}$$

Put differently, instead of comparing actual values of the series in the example mentioned above, period-to-period changes are being compared using the following notation:

$$\Delta y_t = \beta \Delta x_t + \Delta e_t \tag{8}$$

The rationale of this procedure is simple: if we compare the differences of two time series x and y, a strong positive correlation implies that period-to-period changes that are above/below the average for x correspond mainly to changes that are above/below the average for y. Plot D in Figure 2 shows the two example random walks. Plot E demonstrates that there is no significant linear relationship between Δx and Δy ($\beta = 0.01$; $t = 0.28$; $p = 0.78$, $R^2 = 0.00$), while plot F shows no sign of autocorrelation of the residuals, which could be formally tested with various approaches, but does not seem necessary here (Becketti 2013: 380–385)[8]. Thus, building models with changes instead of levels helps to overcome the problems generated by non-stationary time series.[9] Since it is the

[8] For language ontogeny, Baixeries et al. (2013) correlate the level of the exponent of Zipf's law and the level of the mean length of utterances. Bentley et al. (2014) use the GNgC to demonstrate that there is a strong correlation between the level of an economic misery index (which is the sum of the unemployment rate and the inflation rate) and the rate of what they call "literary misery" index (approximating literary mood) both for English and for German. I believe both studies would benefit from a formal test to find out if the involved variables contain a unit root. Indeed, Baixeries et al. (2013: 9) seem to point out that this possibility "should be investigated".

[9] One caveat seems to be in order here: throughout this paper, I implicitly assume that there is no long-term cointegrating relationship between the involved variables. Roughly speaking, cointegration between two time series means that there is a long run equilibrium between the

main goal of the paper to estimate the relationship between different time series, this strategy is better suited than the aforementioned alternative (fitting a dynamic model), because the resulting period-to-period series can directly be used to calculate correlations and partial correlations between the individual series (cf. Section 5).

For the analysis of linguistic and cultural change on the basis of changing word frequency distributions, this means that just because the time series of two words look similar, this does not mean that those words (or n-grams) are related in any substantial sense. However, this special property of time series also offers some interesting novel possibilities for diachronic corpus linguistics: to analyze a potentially evolving relationship between two linguistic structures s_1 and s_2, one can compute a moving-window correlation coefficient between period-to-period changes of s_1 and period-to-period changes of s_2 using the *MVCORR* module of Baum and Cox (2005). Figure 3 presents a couple of examples in this direction based on the American-English GNgC. The left side of each plot calculates the moving-window correlation with a window size of 41. This means for each year t_0, a separate correlation coefficient is calculated that includes all values t_{-20}, t_{-19}, ..., t_{-1}, t_0, t_{+1}, ..., t_{+19}, t_{20}. The right side of each plot shows the two time series for which the correlation is calculated. All data is smoothed with a symmetric five-year window moving-average smoother to highlight the central tendency of the series at each point in time.

Plot A of Figure 3 shows that the unigram “Germany” and the unigram “war” have been positively correlated since the beginning of the twentieth century. This means that a positive change in frequency from one year to another for “Germany” corresponds with a positive change of “war” and vice versa. However, the correlation strength slowly declines in time, potentially indicating that in the present, books that mention “Germany” are not as associated with conflict-laden topics as they once used to be. Plot B in Figure 3 correlates two versions of the present tense of the verb “to be”: the full version “are” and its contraction “(’)re”. The plot reveals that most of the time, year-to-year changes of “are” are negatively correlated with year-to-year changes of “(’) re”, pointing toward a mutual exclusiveness. Maybe a deeper analysis of the

two series which serves as some kind of error correcting device so that if one series changes, at least one of the involved series changes in the future to reestablish balance between the two series (cf. Murray 1994, for a humorous introduction of the concept). While conintegrating relationships are not uncommon for econometric indicators, it is rather unlikely that there is a long-term relationship between the variables measured in this paper. However, this does not imply that it is not an interesting avenue for future research, but the mathematics of this concept (Becketti 2013, S. 385–422) are quite sophisticated and beyond the scope of this paper (and this author).

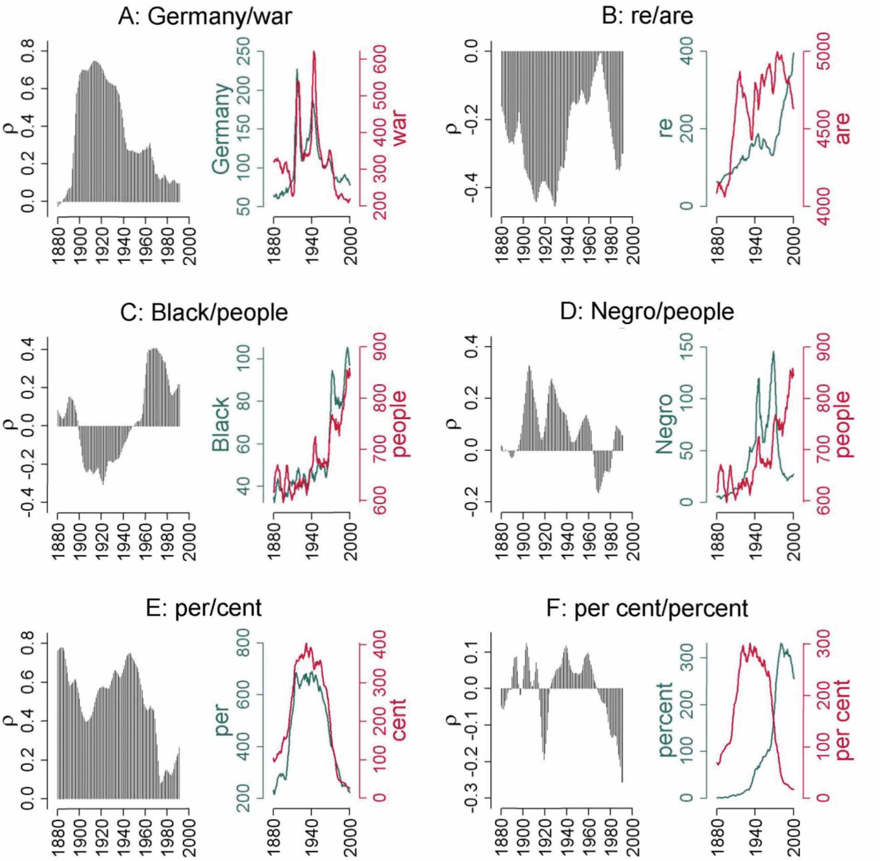


Figure 3: Moving-window correlations coefficients (left side) for the n-grams plotted on the right side (window size: 41 years). Plot A : "Germany" and the unigram "war" were positively correlated since the beginning of the 20th century. The correlation strength slowly declines in time, indicating that in the present, books that mention "Germany" are not as associated with conflict-laden topics as they once used to be. Plot B: "are" and its contraction "(')re". The plot reveals that most of the time, year-to-year changes of "are" are negatively correlated with year-to-year changes of "(')re", indicating a kind of mutual exclusiveness. Plot C and plot D: A positive association of "Black" and "people" since the 1960s, probably as a result of the American Civil Rights movement. Almost at the same time, "Negro" and "people" became negatively associated. Plot E: Analysis of the collocation "per cent". Plot F: Analysis of the 2-gram "per cent" which is replaced by "percent" in the second half of the twentieth century. Please note that the scaling of the ordinates differs for each plot.

timespan from 1960 to 1980, where the correlation was smaller, would help to gain further insights into this relationship. Plot C and plot D in Figure 3 demonstrate that this kind of analysis could also be used to analyze cultural change: a positive association of “Black” and “people” has evolved since the 1960s, probably as a result of the American Civil Rights movement. Almost at the same time, “Negro” and “people” became negatively associated. Plot E in Figure 3 demonstrates that this approach can also be used for the analysis of collocations, such as “per” and “cent”. Finally, the analysis is not restricted to the analysis of 1-grams, but can also be used to compare different linguistic structures, as plot F demonstrates, in which the 2-gram “per cent” seems to be replaced by “percent” after 1970.

4 Variables considered in this study

4.1 The parameters of the Zipf law and the Zipf–Mandelbrot law

Mathematically, Zipf's law can be modeled as a right-truncated zeta distribution (Baixeries et al. 2013), where the probability p of a word with rank r is:

$$p(r) = \frac{1}{\sum_{r=1}^{N} r^{-\alpha}} r^{-\alpha} \qquad [9]$$

Here, N is the observed number of word types, that is, the vocabulary size in a given sample. Correspondingly, the Zipf–Mandelbrot law can be modeled as:

$$p(r) = \frac{1}{\sum_{r=1}^{N} (r+\beta)^{-\alpha}} (r+\beta)^{-\alpha} \qquad [10]$$

Goldstein et al. (2004) show that a linear estimation of the exponent α by log-log transforming the data tends to produce severely biased results. Therefore, a program that evaluates the log-likelihood function as described by Baixeries et al. (2013) was implemented using Stata's *ml* command (Koplenig 2014). A derivation of the log likelihood can be found in Appendix A.2. The maximum number of ML iterations to converge was set to 1,000.[10]

[10] Using R to fit the parameters of the Zipf distribution and of the Zipf–Mandelbrot distribution with the *tolerance* package (Young 2010) yielded identical results compared to the corresponding Stata *zipffit* routine ($\rho = 1.00$ for all pairwise correlations between the estimation with Stata and the estimation with R for all investigated languages).

Throughout this paper, all ranks are used to fit the Zipf and ZM parameters as in the studies of Bentz and colleagues (Bentz et al. 2014a, Bentz et al. 2014b). Beside the actual parameters, Baixeries et al. (2013) also fit a maximum rank and Corral et al. fit (2014) a low frequency cut-off. From a statistical point of view, this is correct, as Clauset et al. (2009) argue. This is due to the fact that for most empirical phenomena, the power law behavior of the data only holds for values greater than some minimum value; the rest of the distribution does not follow a power law. Therefore Clauset et al. (2009) develop a framework based on the Kolmogorov–Smirnov statistic in which, aside from the α parameter, the minimum value x_{min} for which the power-law behavior actually holds, is fitted as well. Figures 10 and 11 in Appendix A.3 present the results using this approach. The results indicate that this approach is not fruitful for the present analysis. This most likely has to do with the fact that in the case of word frequency distributions, a very large number of word types (mostly content words) occur only very rarely. Approximately half of the vocabulary only occurs once in a corpus (Baroni 2009). Therefore, a great deal of the lexical diversity is represented by words that occur only very rarely. Correspondingly, additional analyses reveal that the general tendency of the results presented in Section 6 remains rather stable when word types with a frequency of less than two are excluded from the analysis (cf. Figure 12). When word types with a frequency of less than ten are excluded, the partial correlation of the ZM exponent with the vocabulary size is greatly reduced for all investigated languages while the correlation with the noun–pronoun ratio remains almost unaffected (cf. Figure 13). Nevertheless, it is important to bear in mind that – from a statistical point of view – just because a data set might look as if it follows a power law, this kind of distribution might not be a good (or even a correct) description of the investigated empirical phenomena. How this affects the linguistic interpretation of Zipf curves might be an interesting avenue for future research.

4.2 Lexical richness: vocabulary size

Since the corpus size for each year is held constant at 1,000,000, the vocabulary size v, which is the number of different word types in a given year, can be used as measure of lexical richness.[11] Another widely used measure of lexical richness is the type-token ratio, which is the number of different words (types) divided by

[11] It is noteworthy that due to legal reasons, n-grams that occur less than 40 times in the corpus as a whole are excluded from the GNgC (Michel et al. 2010b). Therefore, the true vocabulary size cannot be calculated.

the total number of words (token) in a text. Because the number of tokens is constant, the type-token ratio and v are identical.[12]

4.3 Syntactic complexity: mean sentence length

For the evolution of children's language, Baixeries et al. (2013) show that the mean length of utterances tends to increase as the exponent of the Zipf distribution decreases. In this context, one can ask if this relationship also holds in a diachronic corpus study.

Both Szmrecsânyi (2004) and Wasow (1997) show that the mean sentence length, that is, the average number of words per sentence can be used as a measure of syntactic complexity. The rationale of this measure is clear: a short sentence will – on average – be syntactically less complex than a longer sentence. Using the mean sentence length (*msl*) as a proxy for syntactic complexity has one considerable advantage compared to other related measures, e.g. counting nodes (of a syntax tree): it is straightforward and easy to operationalize. Westin (2002: 79–81), for example, shows that from 1900 to 1993 the average sentence length in English upmarket editorials decreased from 31.4 to 20.9 words. Therefore she concludes that the sentence complexity decreased during the twentieth century.

For each language, the GNgC data also contain n-grams that indicated sentence beginnings and endings for n-grams with $n > 1$, but no sentence external n-grams. In the following toy example, the two following sentences would be recorded and part-of-speech (*POS*) tagged in the GNgC data using the universal tagset described in Lin et al. (2012) as shown in Table 1: "Hello world! This is an example." This in turn makes use of the information in the first column (or in the second column) to calculate the mean sentence length (*msl*) by dividing the total number of all tokens in a particular year (excluding tokens tagged as punctuation or tokens tagged as indicating the beginning of sentence) by the total number of all tokens tagged as indicating the beginning of a sentence.

[12] For the estimation of the parameter of the Zipf distribution and the parameters of the Zipf–Mandelbrot distribution and for the calculations of the vocabulary size, all upper case strings were converted to lowercase strings and the POS information was removed (i.e. two identical word strings with different POS tags were treated as identical). Re-running the analyses presented in this paper without removing the POS information does not alter the results significantly.

Table 1: Toy example of the 2-gram based representation of the two sentences "Hello world! This is an example." in the GNgC data.

WORD 1	POS 1	WORD 2	POS 2
START	START	Hello	NOUN
Hello	NOUN	world	NOUN
world	NOUN	!	PUNCT
!	PUNCT	_END_	END
START	START	This	PRON
This	PRON	is	VERB
is	VERB	an	DET
an	DET	example	NOUN
example	NOUN	.	PUNCT
.	PUNCT	_END_	END

The following toy example illustrates this procedure: we have ten tokens in total, two tokens tagged as punctuation and two tokens tagged indicating the beginning of a sentence. Therefore, the mean sentence length *msl* would be: (10-2-2)/2 = 3, which is true since the first sentence has a length of two tokens and the second sentence has a length of four tokens.[13]

4.4 Stylistic tendencies: noun–pronoun ratio

Säily et al. (2011) show that noun and pronoun ratios can be used to measure changing stylistic tendencies. Following Biber and Finegan (1989), "pronouny" texts are indicative of an "involved" style, while a more informational style can be characterized by high frequencies of nouns. The analysis of the Longman Corpus of Biber et al. (1999: 65/92; see also Säily et al. 2011) reveals that the distribution of both content words (including nouns) and function words

[13] Unfortunately, it is not completely clear what "sentence external n-gram" actually means. Lin et al. (2012: 77) only write that they applied "a set of manually devised rules" for sentence boundary detection. To account for this fact, I calculated a modified version of the average sentence length, where the sum of all tokens in a particular year (excluding tokens tagged as punctuation or tokens tagged as indicating the end of sentence) is divided by the sum of all bigram-tokens where the first word is tagged as punctuation and is either a full stop, an exclamation mark or a question mark, and where the second word is tagged as indicating the end of sentence. A strong positive correlation of the first differences of the resulting time series for all investigated languages confirms that both approaches lead to similar results ($\rho_{\mathrm{US}} = 0.96$; $\rho_{\mathrm{UK}} = 0.97$; $\rho_{\mathrm{FRE}} = 0.92$; $\rho_{\mathrm{GER}} = 0.91$; $\rho_{\mathrm{ITA}} = 0.99$; $\rho_{\mathrm{SPA}} = 0.94$).

(including pronouns) is indicative of different genres. For example, fiction contains fewer nouns and more pronouns than academic prose.[14]

In the present study, I use the ratio of the proportion of noun and pronouns, that is, the total number of all tokens tagged as nouns (w_{noun}) divided by the total number of all tokens tagged as pronoun (w_{pron}), which can be defined as:

$$r_{np} = \frac{\sum^{f}(w_{\mathrm{noun}})}{\sum^{f}(w_{\mathrm{pron}})} \qquad [11]$$

where v is the vocabulary size and $f(w)$ is the token frequency of word w_{i}.

Again, it is important to emphasize that without metadata, diachronic changes of r_{np} can both approximate (a) a language external effect or (b) a development on a more general level. (a) has to do with the likely fact that the types of books that are in the GNgC are changing as a function of time, while (b) would document a diachronic change across different genres (for an analysis of this differentiation, see Szmrecsanyi 2014).

While this problem does not affect the results presented in this study, it is worth pointing out that Mair et al. (2002) present evidence for (b): they use two one million token corpora of standard written English to show that the proportion of nouns tends to increase over time across different genres, which is accompanied by a decrease of the proportion of pronouns. In a similar vein, Westin (2002: Ch. 3.4/ Ch. 4.2) demonstrates that the frequency of nouns increased from 1900 to 1993 in three different (British-)English newspapers editorials. In the same time span, the use of first person pronouns and the pronoun *it* decreased over time, while the proportion of demonstrative pronouns, second person pronouns and indefinite pronouns remains rather stable.

For all investigated languages, year-to-year changes of the number of nouns and year-to-year changes of the number of pronouns are negatively correlated in the GNgC data: moderately for the Spanish and for the British-English data ($\rho_{\mathrm{SPA}} = -0.53$, $\rho_{\mathrm{SPA}} = -0.66$) and rather strongly for the other languages ($\rho_{\mathrm{US}} = -0.74$, $\rho_{\mathrm{FRE}} = -0.71$, $\rho_{\mathrm{GER}} = -0.73$, $\rho_{\mathrm{ITA}} = -0.82$).

[14] Google also published an English Fiction corpus that predominately consists of fictional works and literary criticism (Michel et al. 2010b). In Appendix A.3/Figure 14 the noun–pronoun ratio is plotted as a function of time for English Fiction, British English and American English. As expected, the English Fiction corpus contains fewer nouns and more pronouns than the other two corpora.

5 Methods

For each investigated language, the three indicators are calculated for each year, from 1800 to 2000.

To induce stationarity for all indicators plus the corpus size, the first differences are taken and tested for a unit root using the Phillips–Perron test (Phillips and Perron 1988). In each case, the Phillips–Perron test showed that taking the first differences of the series results in (weakly) stationary series (all MacKinnon approximate $ps < 0.0001$).

As a next step, the α parameter of the Zipf–Mandelbrot law is compared to the type-token ratio, the noun–pronoun ratio and the mean sentence length. In Appendix A.3 additional results are presented where the following analysis is re-run for both the β-parameter of the ZM law and for the α-parameter of the Zipf law.

To estimate α and calculate v the 1,000,000 random samples are used[15]. The calculation of *msl* is based on the full 2-gram diachronic corpora. It is important to keep in mind that the size of the GNgC is strongly increasing as a function of time. To rule out the potential problem that the procedure of measuring the variables considered in this study is systematically influenced by this increase, year-to-year changes of the respective variables are compared to year-to-year changes of the corpus size. In each case, year-to-year changes are 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 ($|\rho|_{\mathrm{MAX}} = .21$; $|\rho|_{\mathrm{AVERAGE}} = 0.08$). Appendix A.1 shows a table with the resulting values.

The influence of changes of each of the three variables on the changes of the ZM parameter are then interpreted using the coefficient of determination R^2, which is simply the square of the standard Pearson measure ρ. It measures the proportion of the variance of changes of the parameters which is “explained” by changes of each of the three variables. Thus, a high R^2 indicates that knowing the value of the respective variable greatly reduces the error rate when predicting the value of the ZM parameter and vice versa.

In addition to that, partial coefficients of determination are estimated. The partial correlation between the ZM parameter $\Delta\alpha$ and one of the three investigated variables Δvar_1 represents the correlation of $\Delta\alpha$ with Δvar_1 when Δvar_2 and Δvar_3 are held constant. Formally, the partial correlation between $\Delta\alpha$ and Δvar_1

[15] All analyses were case insensitive.

is the correlation between the residuals of a regression of $\Delta\alpha$ on Δvar_2 and Δvar_3 and the residuals of a regression of Δvar_1 on Δvar_2 and Δvar_3.[16]

6 Results

Table 2 summarizes the pairwise correlations for the three variables investigated in this study for all six languages. It is rather unsurprising that the variables are somewhat correlated. The observed correlation between the mean sentence length (measured in word tokens per sentence) and both the vocabulary size and the noun–pronoun ratio is not unexpected, since (at least some) pronouns serve as "economic devices" (Biber et al. 1999: 70) that compete with the full noun phrase (which consists of more words) for the same syntactic position (Mair et al. 2002: 4). However, especially for German and also for American English, this correlation is very low, while it is quite considerable for Italian.

Table 2: Pairwise correlations of the three indicators (year-to-year changes).

Variable 1	Variable 2	American English	British English	French	German	Italian	Spanish
vocabulary size	mean sentence length	−0.623	−0.500	−0.262	−0.343	−0.695	−0.278
vocabulary size	noun–pronoun ratio	0.300	0.277	0.224	0.491	0.550	0.158
mean sentence length	noun–pronoun ratio	−0.182	−0.289	−0.249	−0.129	−0.509	−0.353

The correlation between the vocabulary size and the noun–pronoun ratio can easily be explained: a bigger vocabulary size (measured in the number of different types) is indicative of a richer lexicon, consisting of more content, especially lexical words, for example nouns. Pronouns on the other hand belong to the closed class of words, of which there are relatively few; thus, more

[16] The significance of each R^2 value is calculated using the following formula:

$$p = 2 * ttail\left(n - 2, \frac{|\rho|\sqrt{n-2}}{\sqrt{1-R^2}}\right) \qquad [12]$$

where n is the number of cases and *ttail* returns the reverse cumulative Student's t distribution (StataCorp 2011: 398). The significance of the partial correlation of $\Delta var1$ with $\Delta\alpha$ is the probability that the coefficient of $\Delta var1$ (resulting from a regression of $\Delta\alpha$ on $\Delta var1$, $\Delta var2$ and $\Delta var3$) is equal to 0.

pronouns are to some extent negatively correlated with the vocabulary size. Again of all investigated languages, the correlation is strongest for Italian.

As mentioned above, partial correlations between the respective distributional parameter and the three investigated variables are calculated for each comparison to account for the observed intra-correlations.

As a general descriptive result, Figure 4 shows that there is a general upward trend for both vocabulary size and noun–pronoun ratio for all investigated languages. If we calculate the ratio of the average value of 1990–2000 to the average value of 1800–1810 for both vocabulary size and noun–pronoun ratio, it is

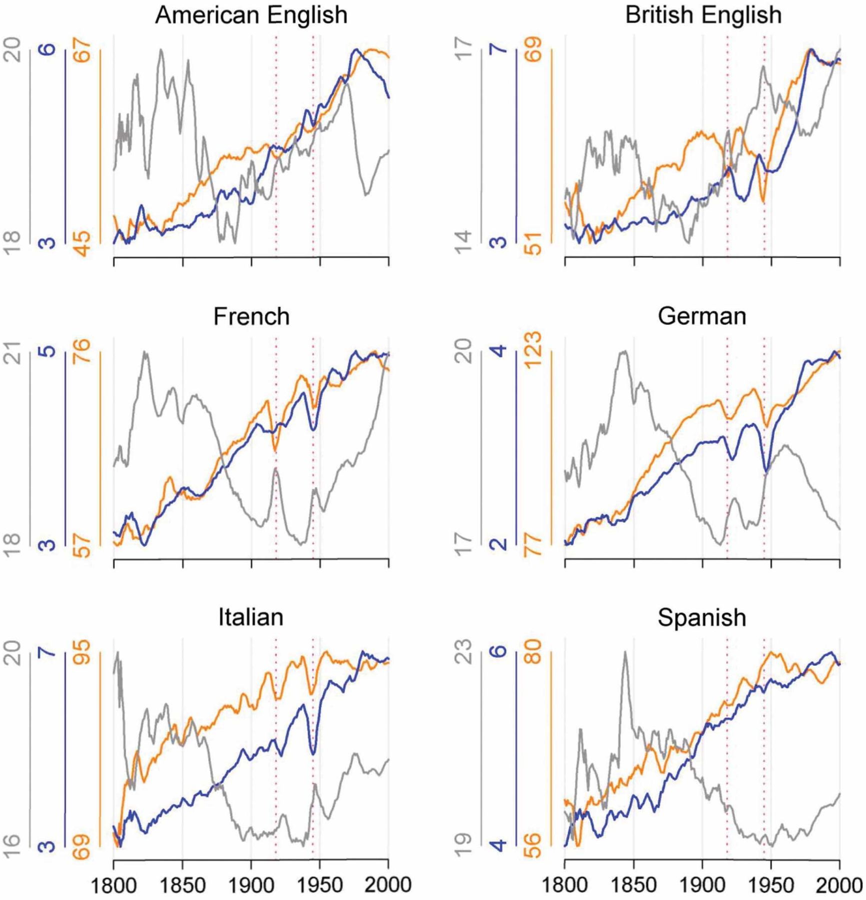


Figure 4: Time series of the vocabulary size (orange, in thousands), the noun–pronoun ratio (blue) and the mean sentence length (gray) for all six investigated languages. The dotted pink lines mark the years 1918 and 1945. All time series smoothed with a symmetric 5-year moving window.

interesting to see that the three Romance languages (French, Italian and Spanish) are somewhat similar both in terms of the change of the vocabulary size and in terms of the change of the noun–pronoun ratio. The other three languages (American English, British English and German) share a similar change of the noun–pronoun ratio, but the change of the vocabulary size is much smaller for British English compared to American English and German. For the mean sentence length, the results are rather mixed and there is no clearly visible trend.

Figure 5 visualizes these results. Another striking aspect in Figure 4 is the fact that for all investigated languages (except the Spanish and the American English data) there are obvious spikes in the period of WWII (1939–1945) and in the period of WWI (1914–1918; both indicated by a thin dotted pink line for the years 1918 and 1945).[17]

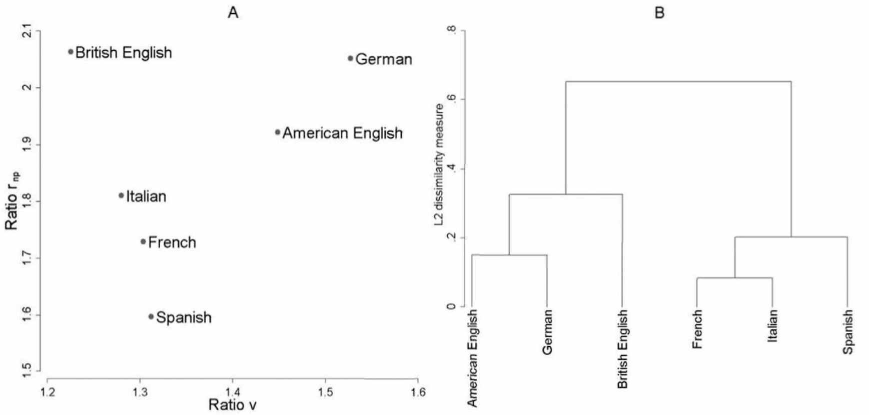


Figure 5: Plot A – the ratio of the average value of 1990–2000 to the average value of 1800–1810 for the noun–pronoun ratio against the vocabulary size. Plot B – dendrogram of a cluster analysis (Ward’s linkage, L2 dissimilarity measure).

Figure 6 shows that there is a general downward trend for both ZM parameters. The spikes in times of WWI and WWII for British English, French, German and Italian are clearly visible. A visual comparison of the time series in Figure 4 and

[17] In Koplenig (submitted), I show that these observations combined with further analyses using an approach put forward by Kilgarriff (2001) point toward a short-term-change in the composition of the GNgC.

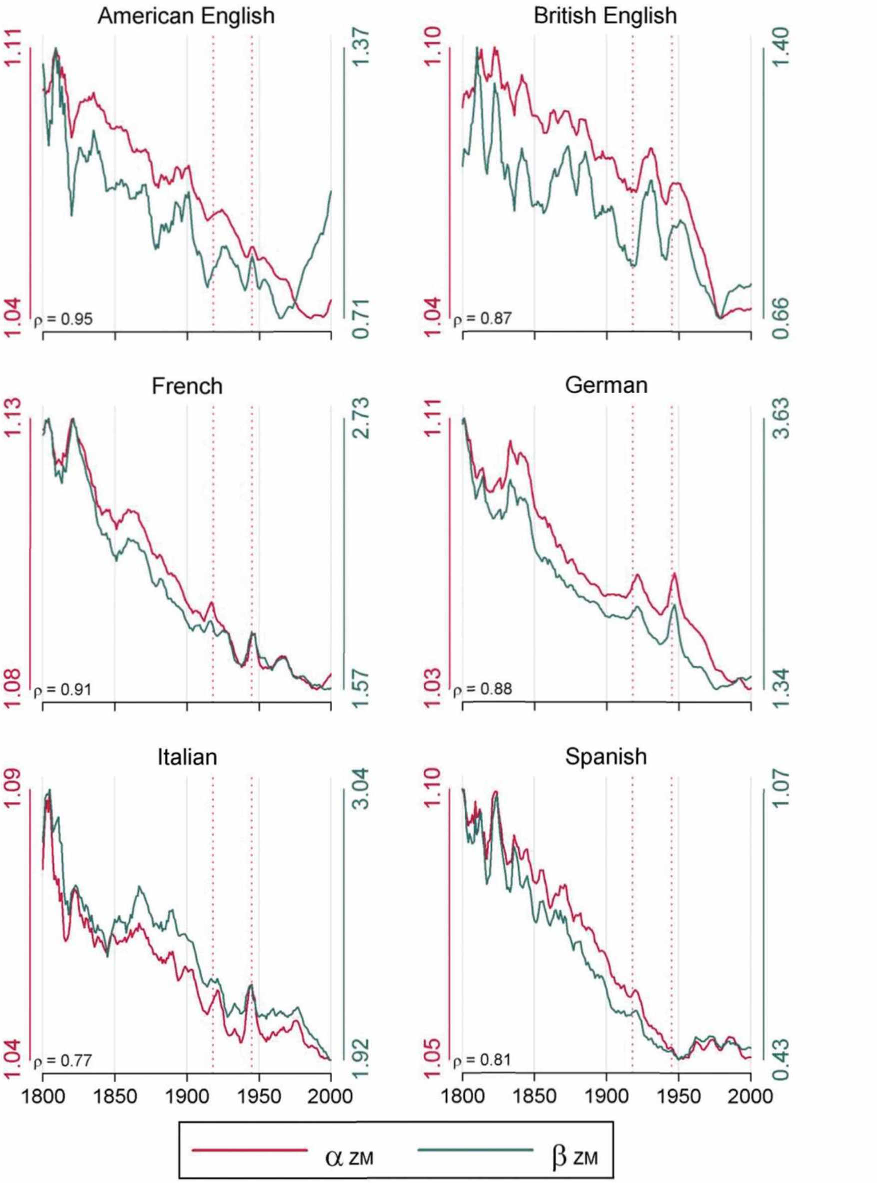
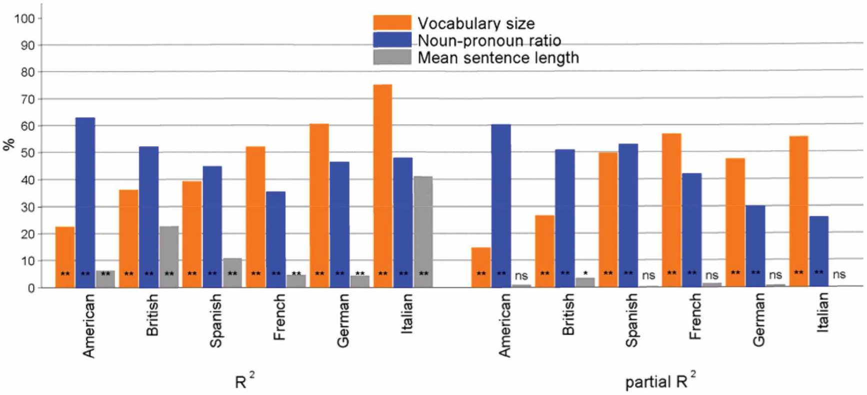


Figure 6: ML estimation of the parameter of the Zipf distribution and the parameter of the Zipf–Mandelbrot distribution as a function of time. Cranberry lines – time series of the Zipf exponent. Emerald lines – time series of the Zipf–Mandelbrot exponent. The dotted pink lines mark the years 1918 and 1945. The ρ-values on the bottom left side of each plot report the correlation values of $\Delta\alpha_{ZM}$ with $\Delta\beta_{ZM}$. All time series smoothed with a symmetric 5-year moving window.

the time series in Figure 6 suggests that the parameters are correlated with the vocabulary size and the noun–pronoun ratio, and to a lesser extent to the mean sentence length.

Since both ZM parameters are strongly correlated and because α_{ZM} is "connected to changes in both low and high frequency words" as Bentz et al. (2014b: 9) note, what follows α_{ZM} is used as an indicator of linguistic change.

The left side of Figure 7 confirms that a great deal of the variance $\Delta\alpha_{ZM}$ can be explained by year-to-year changes of the vocabulary size and changes of the noun–pronoun ratio. Changes of the mean sentence length are only notably correlated with changes of ZM's α for the Italian data and to some extent for the British English data.



ns: not significant / *: p≤0.05 / **: p ≤0.01

Figure 7: Coefficients of determination (left side) and partial coefficients of determination (right side) between year-to-year changes of α_{ZM} and year-to-year changes of the vocabulary size (orange), the noun–pronoun ratio (blue) and the mean sentence length (gray) for all six investigated languages.

However, there are also some important inter-languages differences: while changes of α_{ZM} are correlated more strongly with the noun–pronoun ratio than with changes of the vocabulary size for the two varieties of English, it is the other way round for French, German and Italian (there are no obvious differences for the Spanish data).

The right side of Figure 7 makes this structure clearer: when changes of the vocabulary size (and changes of the mean sentence length) are held constant, then we observe that ZM's α is most responsive to changes of the noun–pronoun ratio for the two English varieties and for Spanish. To a lesser extent, it captures changing noun–pronoun ratios for the French, German and Italian data. For those three languages and for Spanish, changes of ZM's α are more predictive of changes of the vocabulary size when changes of the other two variables are held constant compared to the two English varieties.

It is also noteworthy that when changes of the vocabulary size and changes of the noun–pronoun ratio are held constant, the partial R^2 of the mean sentence length is not significant in all but two cases and does account for less than 10% in the other two remaining cases (the two English varieties).

As an intermediate result, it can be asserted that the Zipf–Mandelbrot distribution captures important aspects of lexical changes and language structure. Table 3 shows that the relationship between changes of the vocabulary size and changes of the noun–pronoun ratio with changes of ZM's α is negative, indicating that when of those two indicators increases, ZM's α "reflects" this with a decrease. This observation is consistent with the results of Bentz et al. (2014a, 2014b) who demonstrate that the Zipf–Mandelbrot law captures linguistic differences in the way simplex and complex concepts are informationally encoded.

Table 3: Pairwise correlations and partial correlations between α_{ZM} and one of three indicators (year-to-year changes).

			Correlation			Partial correlation
Language	**Vocabulary size**	**Noun–pronoun ratio**	**Mean sentence length**	**Vocabulary size**	**Noun–pronoun ratio**	**Mean sentence length**
American English	−0.476	−0.794	0.252	−0.385	−0.777	−0.094
British English	−0.602	−0.722	0.478	−0.516	−0.713	0.182
French	−0.722	−0.594	0.215	−0.754	−0.647	−0.122
German	−0.779	−0.682	0.208	−0.689	−0.547	−0.087
Italian	−0.867	−0.692	0.641	−0.746	−0.509	−0.003
Spanish	−0.627	−0.669	0.330	−0.705	−0.728	−0.052

There is no conclusive evidence that the Zipf–Mandelbrot distribution also measures diachronic changes of the syntactic complexity of language data. However, an alternative explanation of this observation could be that the approach used to measure syntactic complexity in this paper is flawed, for example because the mean sentence length is not a good indicator to measure syntactical complexity in diachronic corpora or there is something wrong with the "set of manually devised rules" (2012: 77) used for sentence boundary detection (cf. Section 4.3). The fact that the correlation between the mean sentence length and the noun–pronoun ratio is somewhat lower than expected potentially points toward this fact (cf. Table 2).[18]

7 Conclusion

The results presented in this paper show that the Zipf–Mandelbrot law can be used to quantify and visualize linguistic change. However, as the additional results indicate, this seems to be true only if the law is fitted to all word types. Or put differently, the estimation procedure does not include the fitting of a threshold token frequency for which the power-law behavior actually holds as suggested by Clauset et al. (2009). It is also important to keep in mind that compared to the vocabulary size as a measure of lexical diversity and compared to the noun–pronoun ratio as a measure of different stylistic tendencies, the parameters of the Zipf–Mandelbrot law do not have an interpretation with an intuitively accessible linguistic content: while it is clear what a vocabulary size of, say, 70,000 in a corpus of 1,000,000 tokens or a noun–pronoun ratio of 3 means, it is not clear without any further point of reference what a corresponding α_{ZM} of 1.11 represents. Thus, finding out that there is a short time increase of α_{ZM} in times of WWI and WWII (cf. Figure 6) for several languages in the GNgC, is somewhat less informative than the observation that the noun–pronoun ratio

[18] To test this explanation, alternative measures of syntactic complexity could be considered (cf. Juola 2008; Ehret and Szmrecsanyi to appear; Montemurro and Zanette 2011; Ramisch 2014; I thank an anonymous reviewer for drawing my attention to those measures). Adopting these approaches to the special structure of the GNgC (only n-grams with $n \leq 5$ that occur at least 40 times in the corpus as a whole) in a computationally feasible manner is an important avenue for future research. Juola's attempt to measure cultural complexity using the GNgC with an information-theoretic approach can be seen as a first step in this direction (Juola 2013). One possible idea could be to adopt the approach used by Piantadosi et al. (2011) and calculate the average amount of information that is used in language comprehension to estimate the amount of cognitive effort required to process a particular word given its preceding context (Frank and Thompson 2012). Instead of randomly deleting words as in the original approach by Juola (2008), syntactic complexity could then be approximated by randomly shuffling a certain amount of the words in the preceding context of the target words.

sharply dropped in this period of time (cf. Figure 4), especially when it comes to hypothesize about the factors causing the observed differences.

On a more general level, the results also show that – with important cross-linguistic differences – a change of the parameters Zipf–Mandelbrot law, describing the shape of a word frequency distribution, can indicate several things such as stylistic changes, lexical changes and probably changes of other linguistic forms[19]. Therefore, it seems safe to say that the Zipf–Mandelbrot distribution can be used as a first indicator to illustrate diachronic change. However, when it is the goal of an investigation "to make previously undetected phenomena available for further analysis and, ultimately, linguistic theory building" (Hilpert and Gries 2009: 398), more thorough analyses should make use of the full spectrum of different lexical, syntactical and stylometric measures to fully understand the factors that actually drive those changes.

Acknowledgments: I thank Julia Kaiser for helping me to check the validity of the corpus cleaning procedure, especially of the French, Italian and Spanish data. I thank my colleague Sascha Wolfer for helping me write an R script to call R from inside Stata and for many helpful discussions regarding the topics presented in this paper. I thank Stefan Engelberg, Carolin Müller-Spitzer, Peter Meyer, Sarah Signer and Sascha Wolfer (again) for (proof-)reading the draft version of this paper and for many helpful discussions. I also would like to thank one anonymous reviewer, whose comments definitely improved this paper. All remaining errors are mine.

[19] An anonymous reviewer pointed out that "when the details of the results are presented and discussed, it is somewhat hard to see a coherent interpretation". I fully agree with her/his impression. However, I do not think that this a drawback of the analyses presented in this paper, but simply reflects the fact that a cross-linguistic "coherent" mapping of the shape parameters of the ZM law to actual trends in language diachrony (and synchrony) is not easy to devise and still requires a lot of empirical research; cf. Bentz et al. (2014b) or Piantadosi (2014) for detailed reviews on the relationship between the form of word frequency distributions and the way languages encode information.

Appendix

A.1 Table of the correlations between all investigated variables and the corpus size (first differences in each case).

Table 4: Correlations of the corpus size with all investigated variables (year-to-year changes) for all investigated languages.

Correlation between corpus size and	British English	American English	French	German	Italian	Spanish
Zipf alpha	0.020	0.072	0.148	0.056	0.129	0.102
Zipf–Mandelbrot alpha	0.142	0.048	0.096	0.063	0.174	0.113
Zipf–Mandelbrot beta	0.173	0.034	0.056	0.060	0.212	0.084
vocabulary size	−0.002	−0.007	0.059	0.028	−0.133	−0.067
mean sentence length	0.025	0.040	−0.005	0.000	−0.005	0.054
noun–pronoun ratio	−0.105	−0.096	−0.099	0.034	−0.045	−0.132

A.2 ML estimation of the parameters of the Zipf law and the Zipf–Mandelbrot law

Since the Zipf law is just a special case of the Zipf–Mandelbrot law (ZM) with $\beta = 0$, the following description focusses on the maximum likelihood fit of the ZM law, while the Stata code presented below includes both options.

In what follows, observations, that is, the word types are assumed to be conditionally independent. Thus, the log-likelihood satisfies the linear form restriction. In Stata, one then only has to specify the log-likelihood function for one individual observation. After that, Stata evaluates this function for every observation and sums up the result. Following Baixeries et al. (2013) the likelihood function for one single word type with rank r and the corresponding frequency fr can be defined as:

$$l_r = p(r)^{f_r} \qquad [13]$$

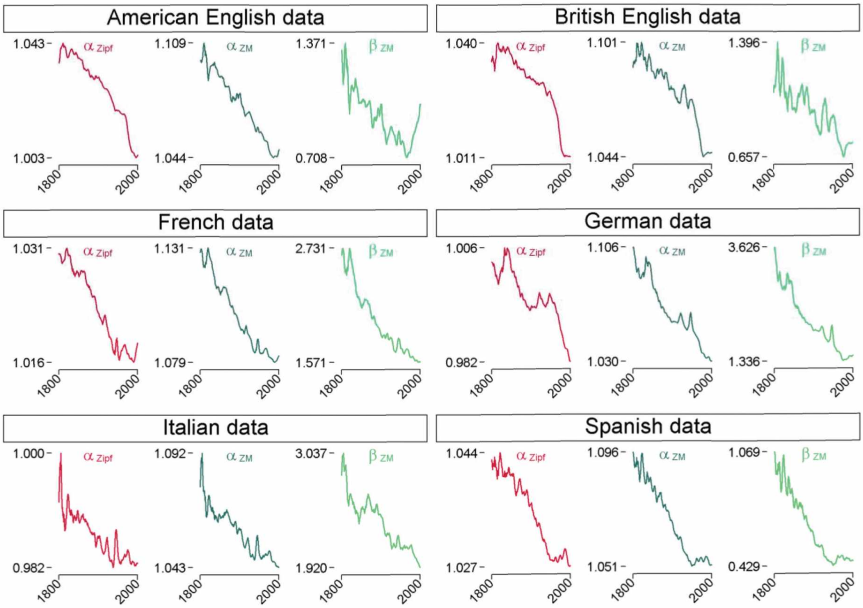
Using the definition presented in eq. [10] and taking logs on both sides this can be rewritten as:

$$\log(l_r) = -\alpha \cdot f_r \cdot \log(r+\beta) - f_r \cdot \log\left(\sum_{r=1}^{N} (r+\beta)^{-\alpha}\right) \qquad [14]$$

A Stata module to fit the one parameter of the Zipf distribution or the two parameters of the Zipf–Mandelbrot distribution by maximum likelihood is available online (Koplenig 2014).

A.3 Additional results

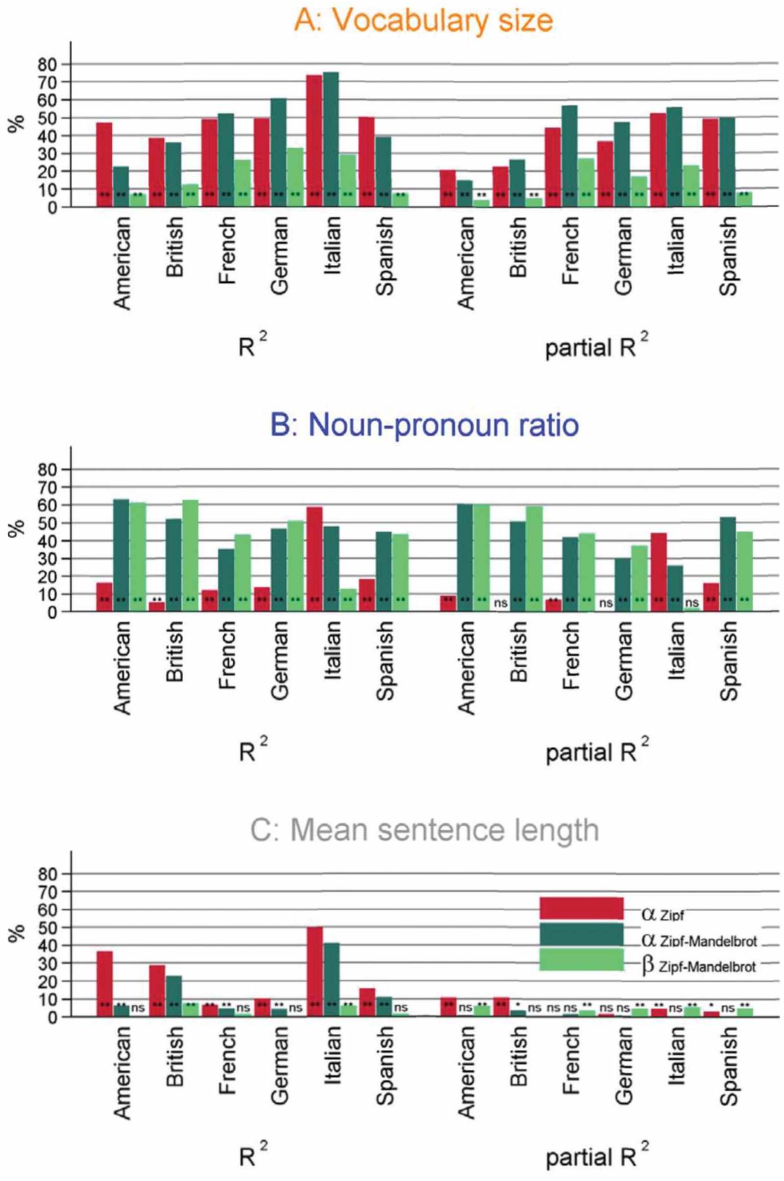
The parameter of the Zipf distribution and the two parameters of the Zipf–Mandelbrot distribution as a function of time.



All time series smoothed with a symmetric 5-year moving window

Figure 8: The parameter of the Zipf distribution (α_{Zipf}) and the two parameters of the Zipf–Mandelbrot (α_{ZM} and β_{ZM}) modification as a function of time.

Correlation-Analysis of the parameter of the Zipf distribution and the two parameters of the Zipf–Mandelbrot distribution with the three indicators.



ns: not significant / *: p≤0.05" / **: p ≤0.01

Figure 9: Coefficients of determination (left side) and partial coefficients of determination (right side) between year-to-year changes of α_{ZIPF} (cranberry), α_{ZM} (emerald), β_{ZM} (mint) and year-to-year changes of the vocabulary size (plot A), the noun–pronoun ratio (plot B) and the mean sentence length (plot C) for all six investigated languages.

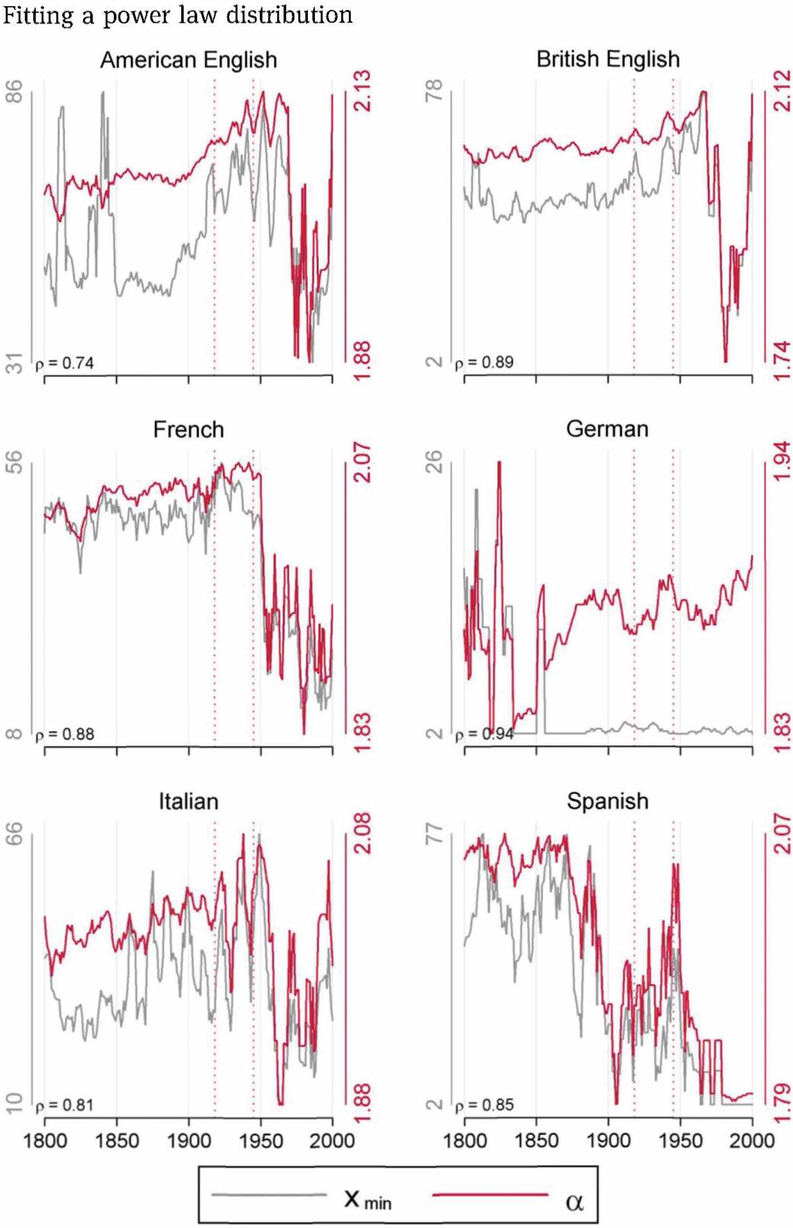


Figure 10: ML estimation of the parameter of a power law as a function time. This analysis used the method presented in Clauset et al. (2007) and the corresponding *plfit* R script developed by Dubroca (2011). Cranberry lines – time series of the α exponent. Emerald lines – time series of the minimum x value. The dotted pink lines mark the years 1918 and 1945. The ρ-values on the bottom left side of each plot report the correlation values of $\Delta\alpha_f$ with Δx_{min}. All time series smoothed with a symmetric 5-year moving window.

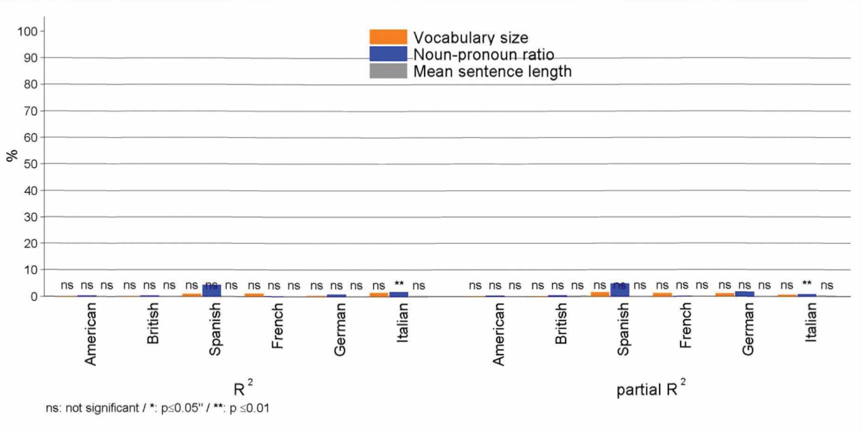


Figure 11: Coefficients of determination (left side) and partial coefficients of determination (right side) between year-to-year changes of power law (using the method presented in Clauset et al. (2007)) exponent and year-to-year changes of the vocabulary size (orange), the noun–pronoun ratio (blue) and the mean sentence length (gray) for all six investigated languages.

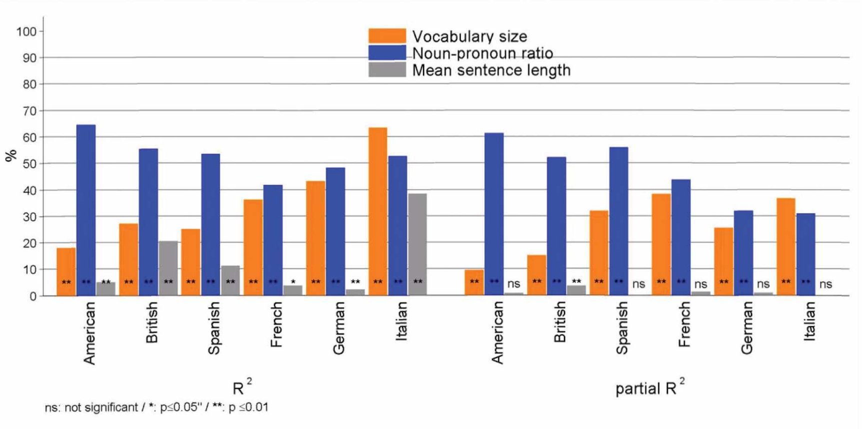


Figure 12: Coefficients of determination (left side) and partial coefficients of determination (right side) between year-to-year changes of α_{ZM} and year-to-year changes of the vocabulary size (orange), the noun–pronoun ratio (blue) and the mean sentence length (gray) for all six investigated languages. Word types with a frequency of less than two were excluded from this analysis.

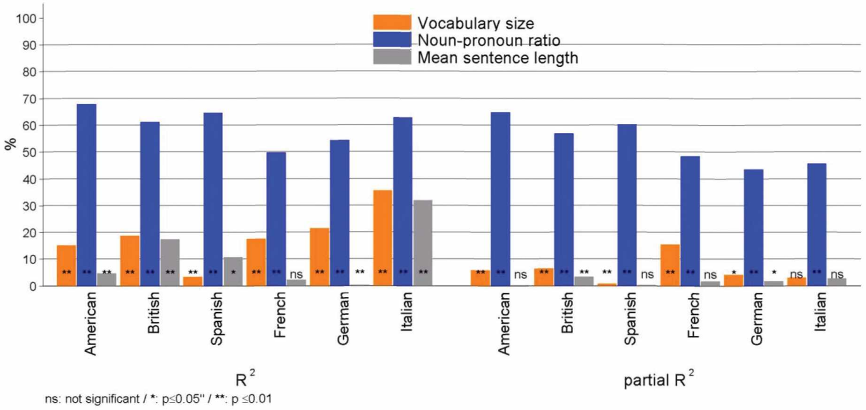


Figure 13: Coefficients of determination (left side) and partial coefficients of determination (right side) between year-to-year changes of α_{ZM} and year-to-year changes of the vocabulary size (orange), the noun–pronoun ratio (blue) and the mean sentence length (gray) for all six investigated languages. Word types with a frequency of less than ten were excluded from this analysis.

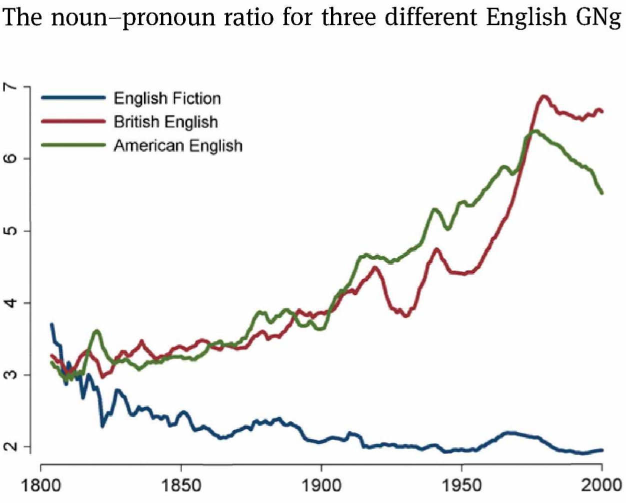


Figure 14: Time series of the noun–pronoun for English Fiction (blue), British English (red) and American English (green). All time series smoothed with a symmetric 5-year moving window.

References

Baixeries, Jaume, Brita Elvevåg & Ramon Ferrer-i-Cancho. 2013. The evolution of the exponent of Zipf's law in language ontogeny. Satoru Hayasaka (ed.). *PLoS ONE* 8(3). e53227. doi:10.1371/journal.pone.0053227 (accessed 10 March 2014).

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 803–821. (Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science Bd. –29.2) Berlin; New York: Walter de Gruyter.

Baum, Christopher F. & Nicholas Cox. 2005. *MVCORR: Stata module to generate moving-window correlation or autocorrelation in time series or panel*. http://ideas.repec.org/c/boc/bocode/s438801.html (accessed 1 September 2014).

Becketti, Sean. 2013. *Introduction to time series using Stata*, 1st edn. College Station, TX: Stata Press.

Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod & Vasileios Lampos. 2014. Books average previous decade of economic misery. Matjaž Perc (ed.). *PLoS ONE* 9(1). e83147. doi:10.1371/journal.pone.0083147 (accessed 10 March 2014).

Bentz, Christian, Douwe Kiela, Felix Hill & Paula Buttery. 2014a. Zipf's law and the grammar of languages: A quantitative study of old and modern English parallel texts. *Corpus Linguistics and Linguistic Theory* 10(2). doi:10.1515/cllt-2014-0009

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paula Buttery. 2014b. Adaptive languages: Modeling the co-evolution of population structure and lexical diversity (submitted). http://www.christianbentz.de/Papers/Bentz%20et%20al.%20(submitted)%20Adaptive%20Languages.pdf (accessed 8 September 2014).

Biber, Douglas. 1991. *Variation across speech and writing*. Cambridge [England]; New York: Cambridge University Press.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487. doi:10.2307/415220 (accessed 1 July 2014).

Biber, Douglas & Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* doi:10.1177/0075424212472509 http://eng.sagepub.com/cgi/doi/10.1177/0075424212472509 (accessed 14 April 2014).

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, England; [New York]: Longman.

Chatfield, Christopher. 2004. *The analysis of time series: An introduction*, 6th edn. (Texts in Statistical Science). Boca Raton, FL: Chapman & Hall/CRC.

Clauset, Aaron, Cosma Rohilla Shalizi & M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4). 661–703. doi:10.1137/070710111 (accessed 10 September 2014).

Clauset, A., M. Young & K. S. Gleditsch. 2007. On the frequency of severe terrorist events. *Journal of Conflict Resolution* 51(1). 58–87. doi:10.1177/0022002706296157 (accessed 10 September 2014).

Corral, Alvaro, Gemma Boleda & Ramon Ferrer-i-Cancho. 2014. Zipf's law for word frequencies: word forms versus lemmas in long texts. http://arxiv.org/abs/1407.8322v1 (accessed 1 October 2014).

Dubroca, Laurent. 2011. *PLFIT*. http://tuvalu.santafe.edu/~aaronc/powerlaws/plfit.r (accessed 12 September 2014).

Ehret, Katharina & Benedikt Szmrecsanyi. 2015 in press. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler & Gudio Seiler (eds.), *Complexity and isolation*, Berlin: De Gruyter. http://www.benszm.net/omnibuslit/EhretSzmrecsanyi_web.pdf (accessed 19 January 2015).

Frank, Stefan L. & Robin L. Thompson. 2012. Early effects of word surprisal on pupil size during reading. In Naomi Miyake, David Peebles & Richard P. Cooper (eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1554–1559. Austin, TX: Cognitive Science Society.

Goldstein, Michel L., Steven A. Morris & Gary G. Yen. 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B* 41(2). 255–258. doi:10.1140/epjb/e2004-00316-5 (accessed 10 April 2015).

Granger, C.W.J. & P. Newbold. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2(2). 111–120. doi:10.1016/0304-4076(74)90034-7 (accessed 23 June 2014).

Hamilton, Lawrence C. 2013. *Statistics with Stata: Updated for version 12*, 8th edn. Boston, MA: Brooks/Cole, Cengage Learning.

Hill, R. Carter. 2008. Principles of econometrics. *Principles of Econometrics*, 3rd edn. *(accompanying website)*. http://www.principlesofeconometrics.com/poe3/poe3do_files/figure12-2.do (accessed 23 June 2014).

Hilpert, M. & S. Th. Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4). 385–401. doi:10.1093/llc/fqn012 (accessed 13 January 2015).

Juola, Patrick. 2008. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change* (Studies in Language Companion Series v. 94) Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Juola, Patrick. 2013. Using the google N-gram corpus to measure cultural complexity. *Literary and Linguistic Computing* 28(4). 668–675. doi:10.1093/llc/fqt017 (accessed 8 April 2014).

Kilgarriff, Adam. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10(2). 135–155.

Kilgarriff, Adam. 2001. Comparing Corpora. *International Journal of Corpus Linguistics* 6(1). 97–133. doi:10.1075/ijcl.6.1.05kil (accessed 19 May 2014).

Koplenig, Alexander. 2015. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram datasets – Reconstructing the composition of the German corpus in times of WWII.

Koplenig, Alexander. 2014. *ZIPFFIT: Stata module to fit the Zipf distribution or the Zipf-Mandelbrot distribution by maximum likelihood*. http://ideas.repec.org/c/boc/bocode/s457872.html (accessed 11 August 2014).

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaar & Khalid Choukri (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)*, 1848–1854. Valetta, Malta: European Language Resources Association (ELRA).

Labov, William. 1994. *Principles of linguistic change.* (Language in Society 20) Oxford, UK ; Cambridge [Mass]: Blackwell.

Lin, Yuri, Jean-Baptiste Michel, Lieberman Erez Aiden, Jon Orwant, Will Brockmann & Slav Petrov. 2012. Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of* *the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174. Jeju, Republic of Korea.

MacWhinney, Brian. 2014. *The Childes Project: Tools for Analyzing Talk, Volume II: the Database*. (Tools for Analyzing Talk). London: Routledge Chapman & Hall. http://www.amazon.com/The-Childes-Project-Analyzing-Database/dp/1138003492/ref=tmm_pap_title_0?ie=UTF8&qid = 1403337096&sr = 1-12 (accessed 21 June 2014).

Mair, Christian, Marianne Hundt, Geoffrey N. Leech & Nicholas Smith. 2002. Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7(2). 245–264. doi:10.1075/ijcl.7.2.05mai (accessed 21 July 2014).

Mandelbrot, Benoît. 1953. An informational theory of the statistical structure of language. In Willis Jackson (ed.), *Communication theory*, 468–502. London: Butterworths Scientific Publications.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Verses, Matthew K Gray, The Google Books Team, Joseph P. Pickett, et al. 2010a. Quantitative analysis of culture using millions of digitized books. *Science* 331(14). 176–182. [online pre-print: 1–12] doi:10.1126/science.1199644.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Verses, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2010b. Quantitative analysis of culture using millions of digitized books (Supporting Online Material). *Science* 331(14). doi:10.1126/science.1199644. http://www.sciencemag.org/content/early/2010/12/15/science.1199644/suppl/DC1 (accessed 5 March 2014).

Montemurro, Marcelo A. & Damián H. Zanette. 2011. Universal entropy of word ordering across Linguistic families. Michael Breakspear (ed.). *PLoS ONE* 6(5). e19875. doi:10.1371/journal.pone.0019875 (accessed 19 January 2015).

Murray, Michael P. 1994. A drunk and her dog: An illustration of cointegration and error correction. *The American Statistician* 48(1). 37–39.

Newman, Mej. 2005. Power laws, pareto distributions and Zipf's law. *Contemporary Physics* 46(5). 323–351. doi:10.1080/00107510500052444 (accessed 10 September 2014).

Phillips, Peter C. B. & Pierre Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75(2). 335–346. doi:10.1093/biomet/75.2.335 (accessed 12 May 2014).

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* doi:10.3758/s13423-014-0585-6 http://link.springer.com/10.3758/s13423-014-0585-6 (accessed 2 May 2014).

Piantadosi, S. T., H. Tily & E. Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526–3529. doi:10.1073/pnas.1012551108 (accessed 21 January 2015).

Ramisch, Carlos. 2014. *Multiword expressions acquisition: A generic and open framework*. New York: Springer.

Säily, Tanja, Terttu Nevalainen & Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2). 167–188. doi:10.1093/llc/fqr004 (accessed 1 July 2014).

StataCorp. 2011. *Stata multivariate statistics reference manual*. Release 12 College Station, TX: StataCorp LP.

Szmrecsanyi, Benedikt. 2004. On operationalizing syntactic complexity. In Gérard Purnelle, Cédrick Fairon & Anne Dister (eds.), *Le poids des mots. Proceedings of the 7th International*

Conference on Textual Data Statistical Analysis 2. 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain.

Szmrecsanyi, Benedikt. 2014. About text frequencies in historical linguistics: disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory*. http://www.benszm.net/omnibuslit/Szmrecsanyi_CH_web.pdf (accessed 8 September 2014).

Tweedie, Fiona J. & R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9(01). 81. doi:10.1017/S0954394500001800 (accessed 29 June 2014).

Westin, Ingrid. 2002. *Language change in English newspaper editorials*. Amsterdam; New York, NY: Rodopi.

Yang, Charles. 2013. Ontogeny and phylogeny of language. *PNAS* 110(16). 6324–6327. http://www.pnas.org/content/early/2013/03/27/1216803110 (accessed 21 June 2014).

Young, Derek S. 2010. Tolerance: An R package for estimating tolerance intervals. *Journal of Statistical Software* 36(5). 1–39.

Zipf, George Kingsley. 1935. *The psycho-biology of language ; an introduction to dynamic philology*. Boston: Houghton Mifflin company.

Zipf, George Kingsley. 2012. *Human behavior and the principle of least effort: an introduction to human ecology*. Mansfield Centre, CT: Martino Pub.