Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions

Alexander Koplenig

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

Abstract

Recently, a claim was made, on the basis of the German Google Books 1-gram corpus (Michel *et al.*, Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 2010; **331**: 176–82), that there was a linear relationship between six non-technical non-Nazi words and three 'explicitly Nazi words' in times of World War II (Caruana-Galizia. 2015. Politics and the German language: Testing Orwell's hypothesis using the Google N-Gram corpus. *Digital Scholarship in the Humanities* [Online]. http://dsh.oxfordjournals.org/cgi/doi/10.1093/llc/fqv011 (accessed 15 April 2015)). Here, I try to show that apparent relationships like this are the result of misspecified models that do not take into account the temporal aspect of time-series data. The main point of this article is to demonstrate why such analyses run the risk of incorrect statistical inference, where potential effects are both meaningless and can potentially lead to wrong conclusions.

Correspondence:
Alexander Koplenig,
Institute for the German language (IDS)
Postfach 10 16 21, 68016
Mannheim, Germany.
E-mail:
koplenig@ids-mannheim.de

1 Introduction

'It is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly "significant." As the occurrence of such "nonsense-correlations" makes one mistrust the serious arguments that are sometimes put forward on the basis of correlations between time-series […] it is important to clear up the problem how they arise and in what special cases.'

(Yule, 1926, p. 2)

'So-called univariate time-series analysis actually is the analysis of the bivariate relationship between the variable of interest and time.'

(Becketti, 2013, p. 92).

The idea to quantitatively study 'the relationship between political regimes and language' (Caruana-Galizia, 2015, p. 1) is certainly a highly interesting

research topic, which became possible with the recent availability of large machine-readable diachronic corpora such as the COHA (Davies, 2010) or the Google N-gram corpora (Michel et al., 2010). The latter, in particular, received widespread attention, as it reportedly contains roughly 4% in the 2009 version (Michel et al., 2010) and even 6% in the 2012 version of all books ever published (Lin et al., 2012). For example, Petersen et al. (2012, p. 4) reason that observable frequency effects in the Google Books N-gram corpora 'during WWII represents a "globalization" effect, whereby societies are brought together by a common event and a unified media', while Bochkarev et al. (2014, p. 1) argue for a '[m]ajor societal transformation'. In a similar vein, Michel et al. (2010) try to demonstrate that censorship in those corpora can be detected by measuring changes in the number of times the name of a person is mentioned. This assumption is tempting, but can be contested, because the Google Books data sets are not accompanied by any metadata regarding the books the corpora consist of, as I try to show in (Koplenig, 2015b [to appear]).

In a recent paper, Caruana-Galizia (2015) uses the German Google Books N-gram corpus to show that there was a linear relationship between six non-technical non-Nazi words and three 'explicitly Nazi words' in times of World War II. This relationship is used as evidence for a hypothesis made by George Orwell 'that everyday language deteriorates under dictatorships' (Caruana-Galizia, 2015, p. 1).

In this article, I first replicate this result (Section 2). I then try to demonstrate why such analyses that do not take into account the special nature of time-series data, run the risk of incorrect statistical inference, where potential effects are both meaningless and can potentially lead to wrong conclusions (Section 3). When one accounts for this problem, the claimed relationship almost disappears entirely (Section 4). This article ends with some concluding remarks (Section 5). To ensure maximal replicability, the Appendix contains a Stata script ('do-file') that automatically downloads the data and reproduces all results presented in this article.

2 Replication of Caruana-Galizia (2015)

To analyze the relationship between six non-technical non-Nazi words (Demokratie [democracy], Freiheit [freedom], Frieden [peace], Herrlichkeit [glory], Gerechtigkeit [justice], and Heldentum [heroism]) and three 'explicitly Nazi words' (Rassenschande [racial defilement], Halbjude [half Jew], and Arier [Aryan]) in times of World War II, Caruana-Galizia (2015) extracts the time-series for each keyword (time span: 1870–1946) from the 2009 version of the German Google Books 1-gram Corpus (Michel et al., 2010).[1]

Figure 1 presents a replication of the (Pearson) correlation analysis that Caruana-Galizia (2015) presents in Table 3. The correlations are comparable for Rassenschande and Halbjude, but are not identical. To make sure that the analyses presented here are correct, I manually extracted each times series from the Google Books 1-gram corpus and the total 1-gram frequency for the time span 1870–1946 and recalculated the correlations with identical differences between my results and the results presented by Caruana-Galizia (2015).[2] The reason for the difference might be that Caruana-Galizia (2015) calculates the overall token 1-gram frequency on the basis of all words that appear in the corpus for each year. Due to legal reasons, however, n-grams that occur less than 40 times in the corpus as a whole are excluded from the Google Books N-gram corpora (Michel et al. 2010b) but are available in the total counts file. Nevertheless, this potential difference cannot explain the huge difference between the Caruana-Galizia (2015) results and the results presented here for the keyword Arier. For example, Caruana-Galizia (2015) finds a correlation between Arier and Herrlichkeit of $r = 0.33$. In my analysis, this correlation is virtually nonexistent ($r = 0.07$).

While it is hard to speculate about potential reasons for this difference, the main problem of an analysis, such as the one Caruana-Galizia (2015) conducts, is the fact that it does not take the special nature of temporally ordered data into account.[3] In the next section, I will outline the problem and explain why it also matters in the analysis I replicated here.
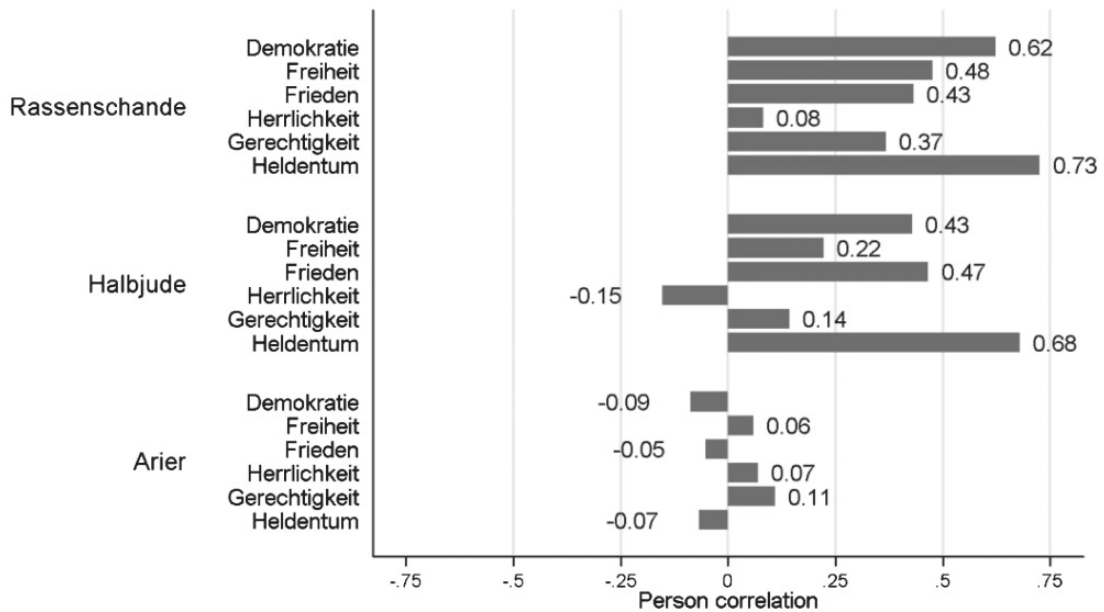
Fig. 1. Replications of the correlation analysis of Caruana-Galizia (2015, p. 11 Table 3). The figure shows Pearson correlations between the Nazi words and the selected keywords in the time span 1870–1946 on the basis of the Google Books German 1-gram corpus (version 2009).

To demonstrate why I believe that the validity of such an analysis can be questioned, I use three additional time-series. The first two time-series are the frequency profiles of two word types related to Switzerland. In Koplenig (2015b [to appear]), I adapted a method for the measurement of synchronic corpus (dis-)similarity put forward by Kilgarriff (2001) to reconstruct the composition of the German corpus in times of World War II. In the absence of information about the texts that the German Google Books corpus compiles, this analysis supports the argument that the corpus was strongly biased toward volumes published in Switzerland during World War II. The two word types that contribute most to the calculated difference are *Zürich* [Zurich] and *Schweiz* [Switzerland]. The frequency profiles of those two words where extracted in the same way as the other keywords.

The third time-series is a simulation of a random walk (henceforth *Randomwalk*) with drift (Hill, 2008; Becketti, 2013, p. 72/73; Koplenig, 2015c), where the value x_t at time point t is given as:

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

with e_t normally distributed in the interval [0,1]. This means that the resulting time-series x has an average upward trend, but otherwise behaves in a completely random manner.

3 The Problem: Pearson Correlation and Non-stationarity

The statistical analysis of time-series—that is, data with a natural temporal ordering—is special. In fact, it is so special that most of the classic statistical tools of data analysis cannot be used directly. In many situations, this has to do with the sequential dependence of observations and with the fact that the variable which is measured at successive moments in time exhibits an upward or downward trend. The resulting series is said to have a unit root or to be non-stationary (Becketti, 2013, pp. 376–85). 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

Table 1. Augmented Dickey-Fuller tests for unit roots. For each word, the test was run for a lag length of 1

Keyword	P-value
Arier	0.00
Halbjude	0.22
Rassenschande	0.89
Demokratie	1.00
Freiheit	0.73
Frieden	0.42
Herrlichkeit	0.26
Gerechtigkeit	0.90
Heldentum	0.88
Zürich	1.00
Schweiz	1.00
Randomwalk	0.69

(Granger and Newbold, 1974; Koplenig, 2015c). There are formal ways to test for unit roots, the classic one is the augmented Dickey-Fuller test (Becketti, 2013, ch. 10.2). I chose this one, since Caruana-Galizia (2015, fn. 10) also used it in later analyses. Table 1 lists the (MacKinnon approximate) P-values for each case.

The null hypothesis states that the respective time-series follows a unit root, or put differently that it evolves through time. For *Arier* the null hypothesis of a unit root can be rejected at $P < 0.01$. However, there is also virtually no correlation for this word and any of the keywords (cf. Fig. 1). For all other keywords including the three additional words, there is good reason to accept the presence of a unit root because the P-value is greater than 0.1. This result points toward the fact that for those words, the time-series seem to be non-stationary.

Why is this problematic? Basically, the Pearson product–moment correlation coefficient is the covariance of two variables x and y scaled to the interval $[-1,1]$. Now, if we have two series x and y that both have an upward trend, then by definition, for both series the following statement is true: values that are later in time will be above average from the mean value of the series, while values that are earlier in time will be below average. Since the covariance measures whether values of x that are above/below average tend to co-occur with values of y that are above/below average, then by

mathematical necessity, the correlation coefficient will be high when in fact they are not related in any substantial sense (Granger and Newbold, 1974). Thus, for two trending time-series, the Pearson correlation only measures the fact that the two series are trending.

Figure 2 presents four plots that all document an apparent linear relationship. To visualize why I believe that the problem described above is also present in the analysis of Caruana-Galizia (2015), the observed values are colored by decade with earlier decades colored in lighter shades of gray and later decades colored in darker shades of gray (as indicated by the color bar at the bottom of the figure). Plot A replicates the findings of Caruana-Galizia (2015) for the Nazi word *Halbjude* and the keyword *Frieden*. At first glance, there seems to be a positive correlation between the time-series of both words as argued by Caruana-Galizia (2015). However, the color pattern reveals that this could be the result of a spurious model: values for later decades (dark shades of gray) strongly influence the apparent relationship. This can be best understood if we have a look at Plot B that shows the relationship between *Randomwalk* and the keyword *Demokratie*. Again, the apparent correlation ($r = 0.66$) is the result of a misspecified model with values for later decades strongly influencing the result. It is noteworthy to point out again that the *Randomwalk* series has an average upward trend, but behaves completely randomly apart from that. So, what other explanation for the observed calculation could be there apart from a spurious model? Plot C shows the relationship between the Nazi word *Rassenschande* and *Zürich*. Again, it is hard to come up with an explanation for this result other than a misspecified model. In Plot D the relationship between *Schweiz* and *Zürich* is depicted. While, as argued below, the time-series of both words are indeed related, the very strong linear relationship ($r = 0.93$) is the result of the fact that both series are trending as indicted by the color pattern.

In the next section, I will outline a procedure to account for this problem and show that this procedure strongly affects the results of an analysis, like the one conducted by Caruana-Galizia (2015).
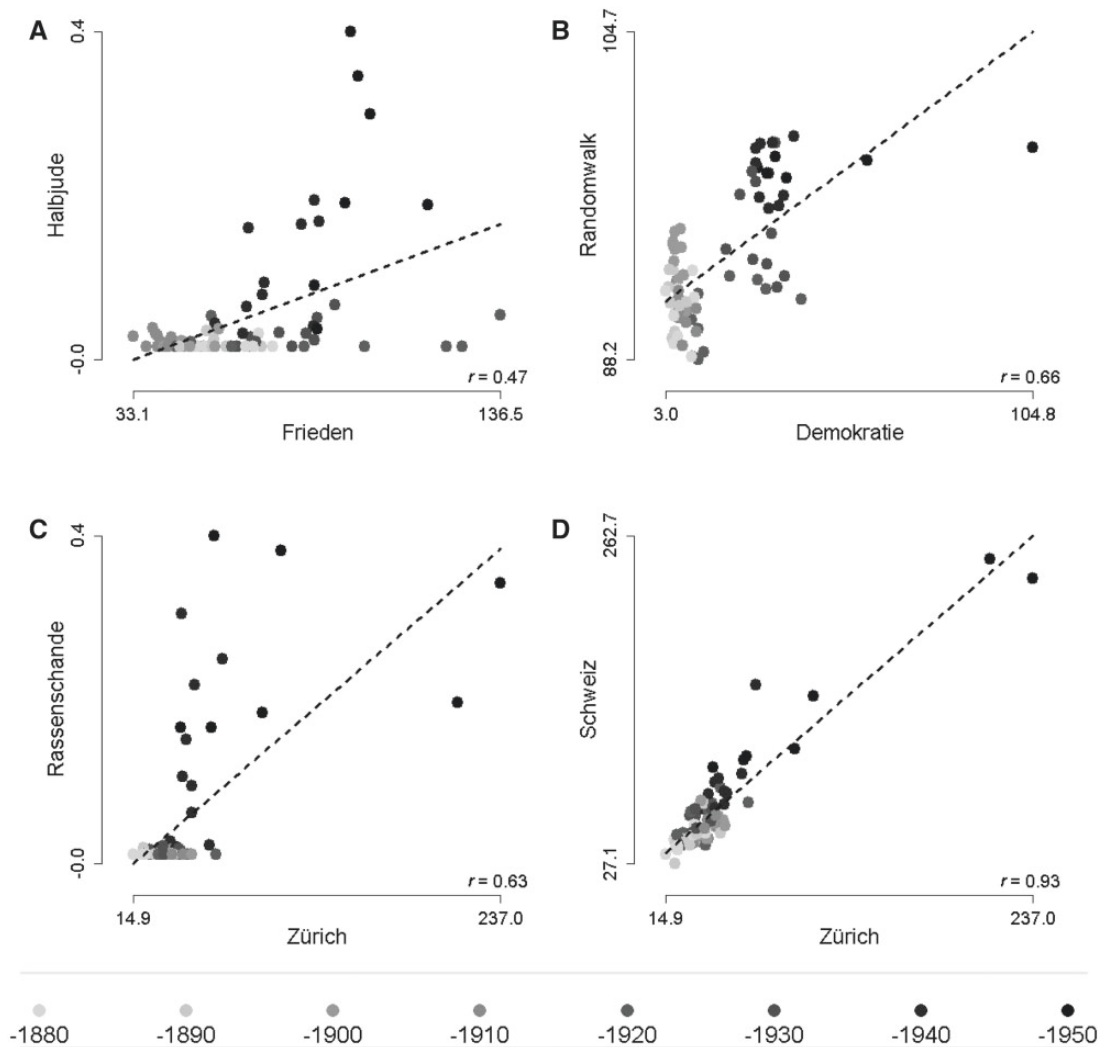
Fig. 2. Linear relationship in the time span 1870–1946 between *Halbjude* and *Frieden* (A), *Randomwalk* and *Demokratie* (B), *Rassenschande* and *Zürich* (C), and *Schweiz* and *Zürich* (D). In each case, a positive linear correlation is found, as indicated by the dashed line (Pearson correlation coefficients are shown in the bottom right corner of each plot). Additionally, the observed values are colored by decade, with earlier decades colored in lighter shades of gray and later decades colored in darker shades of gray (as indicated by the color bar at the bottom of the figure). The fact that there is an obvious color pattern in all four plots (with later decades having most influence on the apparent relationship) supports the claim of a spurious result in each case. *Note*: word frequencies are relative per 1 million words.

4 Accounting for Autocorrelation Questions Apparent Effects

Instead of comparing the actual time-series, one can take the first differences of the variables involved, to induce (weakly) stationarity. Put differently, instead of comparing actual values of the series, period-to-period changes are being correlated. 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.

It is noteworthy that this procedure seems to be better suited to answer a research question like the one Caruana-Galizia (2015) tries to answer: if the relative use of a Nazi word increases from last year to this year, then—on average—the relative use of one of the keywords should also increase from last year to this year if both words are related.

Table 2 demonstrates that this procedure results in weakly stationary series for all keywords, except for *Demokratie.* For this series, it might be appropriate (or necessary) to difference the difference (second-order difference). However, since the correlation analysis presented below shows that even under the assumption of non-stationarity, *Demokratie* does not correlate with any of the three Nazi words beyond random fluctuations, this option is not pursued any further

In Fig. 3, year-to-year changes are correlated instead of actual levels for the selected words. This procedure strongly counters the analysis of Caruana-Galizia (2015): Only *Rassenschande* and *Heldentum* are positively correlated, while most of the correlations are now negative and/or virtually nonexistent. Figure 4 modifies the analysis of Fig. 2 by correlating year-to-year changes. The fact that compared to Fig. 2, the color pattern is less obvious supports the
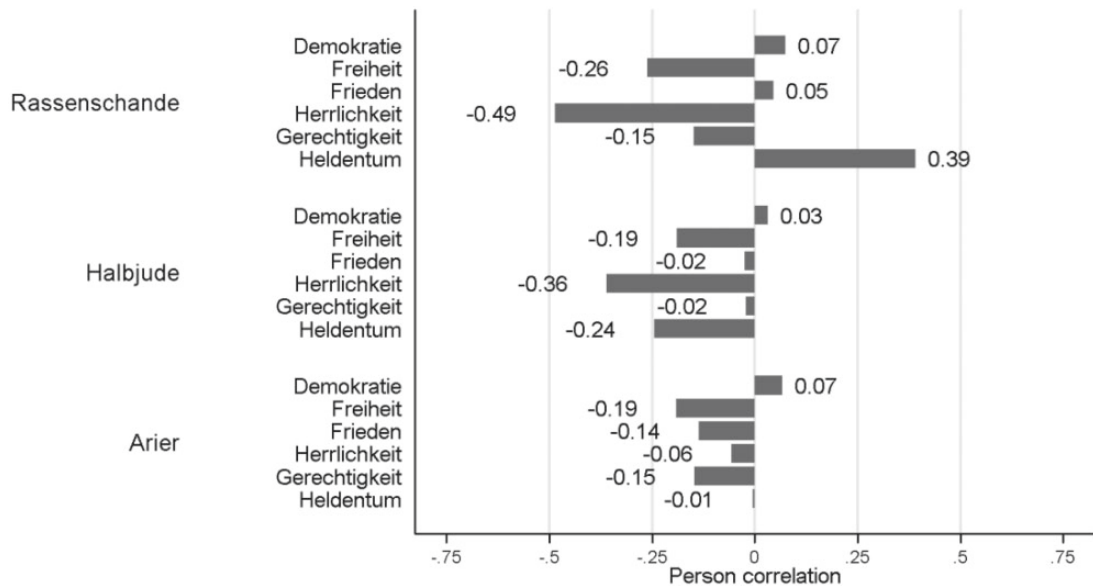
Table 2. Augmented Dickey-Fuller tests for unit roots. For this analysis, the first differences for each time-series were used. For each word, the test was run for a lag length of 1

Keyword	*P*-value
Arier	0.00
Halbjude	0.00
Rassenschande	0.00
Demokratie	0.89
Freiheit	0.00
Frieden	0.00
Herrlichkeit	0.00
Gerechtigkeit	0.00
Heldentum	0.00
Zürich	0.00
Schweiz	0.00
Randomwalk	0.00

Fig. 3. Replications of the correlation analysis of Caruana-Galizia (2015, p .11 Table 3). The figure shows Person correlations between the first differences for the time-series for each word.
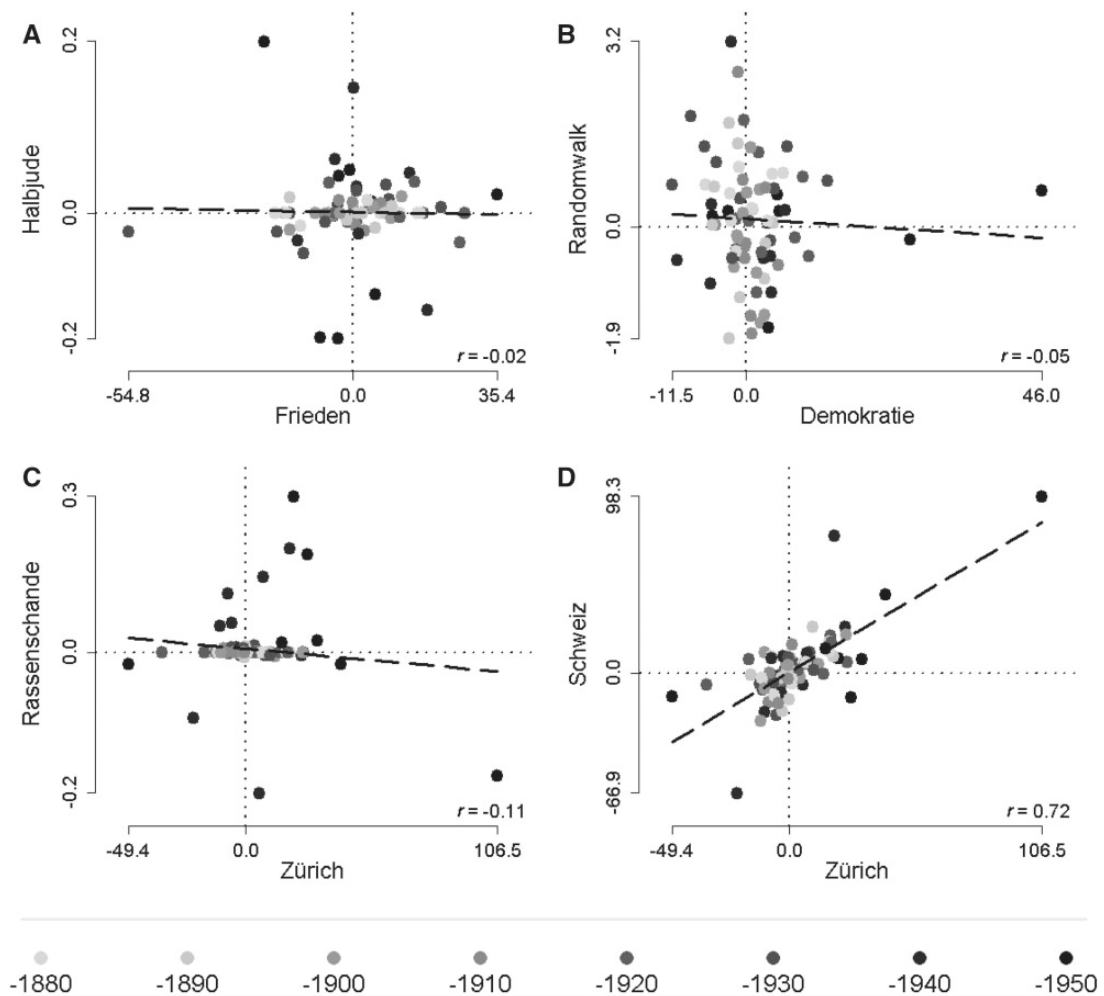
Fig. 4. Linear relationship in the time span 1870–1946 between the first differences for *Halbjude* and *Frieden* (A), *Randomwalk* and *Demokratie* (B), *Rassenschande* and *Zürich* (C), *Schweiz* and *Zürich* (D). The depicted information is described in Fig. 2. The fact that compared to Fig. 2, the color pattern is less obvious supports the claim that the procedure of taking first differences helps to solve the problem of non-stationarity. However, the only noteworthy linear relationship remains between *Schweiz* and *Zürich* (plot D).

claim that the procedure of taking first differences helps to solve the problem of non-stationarity. Correspondingly, there is no linear relationship between year-to-year changes of the *Randomwalk* series and year-to-year changes of the keyword *Demokratie* (plot B). The only noteworthy linear relationship remains between *Schweiz* and *Zürich* (plot D).

On a more general level, I believe that is important not to forget that '[v]isual inspection plays a key role in time-series analysis' (Hamilton, 2013, p. 356; cf. also Becketti, 2013, ch. 11). To this end, Fig. 5 plots the time-series for each combination of words that were presented in Figs 2 and 4. In accordance with Fig. 4, this visual inspection clearly demonstrates that only the time-series for *Schweiz* and *Zürich* seem to behave in a similar way (plot D), while all other plots do not indicate a relationship in any substantial sense.
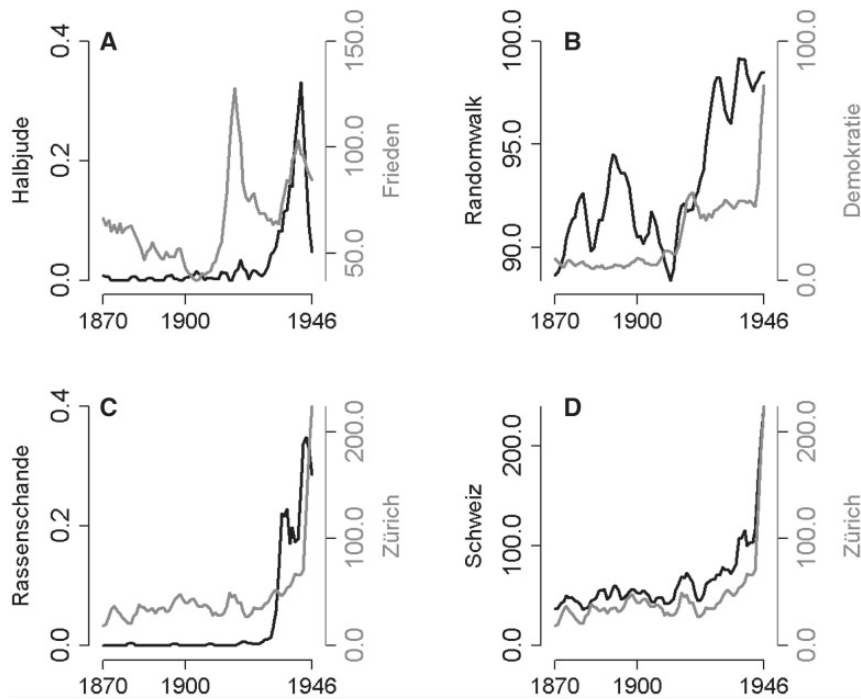
Fig. 5. Time-series plots for the examples presented in Figs 1 and 4. For each plot, the time-series of the first word is placed on the left *y*-axis and colored in black, while the time-series of the second word is placed on the right *y*-axis and colored in gray. Each time-series was smoothed using a simple weighted moving average with a 3-year window centered on the current frequency.

5 Concluding Remarks

The main point of this article was to demonstrate why an analysis of diachronic data that does not take the temporal aspect of time-series data into account, runs the risk of incorrect statistical inference, where potential effects are meaningless and therefore can potentially lead to wrong conclusions. To this end, I replicated the result of Caruana-Galizia (2015, p. 14) who argues that six non-technical non-Nazi words are highly correlated with explicitly Nazi words in order to test a hypothesis by George Orwell, who argues that 'ordinary language deteriorates under dictatorship' (Caruana-Galizia, 2015, p. 14). I hope that the re-analysis presented in this article shows that this result can (or has to) be questioned.[4] In a similar vein, Frimer et al. (2015) claim that there is a linear relationship between the level of prosocial language

and the level of public disapproval of US Congress. Again, a reanalysis casts doubt on this apparent relationship by demonstrating that it is the result of a misspecified model that does not account for first-order autocorrelated disturbances resulting from non-stationarity (Koplenig, 2015a).

Conversely, I believe that the use of more appropriate tools for the analysis of time-series data can help the digital humanities to uncover the 'true' and sometimes potentially even more interesting mechanism of how particular systems or institutions work as I have argued elsewhere (Koplenig, 2015b [to appear]).

Acknowledgments

drafts of this article, Sarah Signer for proofreading, and one anonymous reviewer for her/his valuable comments. All remaining errors are mine.

References

Becketti, S. (2013). *Introduction to Time Series Using Stata*. 1st ed. College Station, TX: Stata Press.

Bochkarev, V., Solovyev, V. and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. [Online]. http://wwwstaff.eva.mpg.de/%7Ewichmann/LexEvolUploaded.pdf (accessed 12 June 2014).

Carmody, S. (2014). *Ngramr: Retrieve and Plot Google N-Gram Data*. [Online]. http://cran.r-project.org/web/packages/ngramr/index.html (accessed 20 April 2015).

Caruana-Galizia, P. (2015). Politics and the German language: Testing Orwell's hypothesis using the Google N-Gram corpus. In: *Digital Scholarship in the Humanities* [Online]. http://dsh.oxfordjournals.org/cgi/doi/10.1093/llc/fqv011 (accessed 15 April 2015).

Davies, M. (2010). The Corpus of Historical American English: 400 million words, 1810–2009. [Online]. http://corpus.byu.edu/coha/ (accessed 16 October 2014).

Frimer, J. A., Aquino, K., Gebauer, J. E., and Zhu, L. (Lei), *et al.* (2015). A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences* 112: 6591–4.

Granger, C. W. J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2: 111–20.

Hamilton, L. C. (2013). *Statistics with Stata: Updated for Version 12*. 8th ed. Boston, MA: Brooks/Cole, Cengage Learning.

Hill, R. C. (2008). *Principles of Econometrics* [Online]. http://www.principlesofeconometrics.com/poe3/poe3do_files/figure12-2.do (accessed 23 June 2014).

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics* 6: 97–133.

Koplenig, A. (2015a). Autocorrelated errors explain the apparent relationship between disapproval of the US Congress and prosocial language. [Online]. http://hdl.handle.net/10932/00-027E-F9B1-E746-3A01-2 (accessed 29 June 2015).

Koplenig, A. (2015b). 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. *Digital Scholarship in the Humanities*, Oxford: Oxford University Press, 2015.

Koplenig, A. (2015c). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory* [Online] 0. http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2014-0049/cllt-2014-0049.xml (accessed 19 April 2015).

Lin, Y., Michel, J. -B., Aiden, L. E., Orwant, J., Brockman, W., and Petrov S. (2012). *Syntactic Annotations for the Google Books Ngram Corpus, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic,* Jeju, Republic of Korea, pp. 169–174.

Michel, J. -B., Shen, Y. K., Aiden, A. P., Verses, A., Gray, M. K., The Google Books Team, **Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J, Pinker, S., Nowak, M. A., and Aiden E. L.** (2010). Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–82. [online pre–print: 1–12].

Petersen, A. M., Tenenbaum, J. N., Havlin, S. and Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* [Online] 2. http://www.nature.com/doifinder/10.1038/srep00313 (accessed 10 March 2014).

Yule, G. U. (1926). Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89: 1–64.

Notes

1 As an aside: the terminology of Caruana-Galizia (2015) is somewhat unclear, in Fig. 1 (2015, p. 8 see also p. 12) he says that the plot shows the '[p]roportion of German books containing keywords, 1870–1946'. This would mean that he uses the relative number of books that contain one of the keywords per year. On page 11, however, he states that 'these correlations show us that when the *relative use* of an explicitly Nazi word increases, so did the keywords' (my emphasis, see also p. 13). This in turn would mean that he uses the relative token frequency of a keyword per year. Both types of information are available in the Google Books corpora; the data sets are freely available here: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (last accessed 28 April 2015). To find out which information Caruana-Galizia (2015) actually uses, I compared the

results he presents in Table 2 with the original data. This shows that he seems to have used 'the relative token frequency'. On this basis, I replicated the analysis. The relative token frequency per year is calculated by dividing the absolute token frequency with the total number of 1-grams. This information is available here: http://storage.googleapis.com/books/ngrams/books/googlebooks-ger-all-totalcounts-20090715.txt (last accessed 04/28/2015).

[2] In addition, a replication of the results with R using the *ngramr* package (Carmody, 2014) yields identical results. I would like to thank my colleague Sascha Wolfer for running this analysis.

[3] This comes as a bit of a surprise since he deals with this problem in further analyses he presents in his article (cf. footnote 4).

[4] Of course, from this does not follow that other results presented in Caruana-Galizia (2015) have to be challenged, too. However, the autoregressive integrated moving average (ARIMA or ARMAX) models he uses in order to predict the relative frequency of a keyword on the basis of the *POLITY2* score (a measure of the level of democracy, the data are available here: http://www.systemicpeace.org/inscrdata.html, last accessed on 28 April 2015) are quite sophisticated and fitting such a model requires several conceptual decisions regarding the appropriate ARIMA structure that depend on each respective time-series (Becketti, 2013, ch. 7). Caruana-Galizia (2015, p. 11) only uses one ARIMA model specification for the time-series of all six keywords. To see why this is rather problematic, I ran separate ARIMAs of *Heldentum* and *Zürich* on the *POLITY2* score (the code that replicates this analysis can also be found in the Appendix). A look at the autocorrelations and partial-autocorrelations of a regression of the first difference of the *Heldentum* series on the first difference of the

POLITY2 score shows that the residuals have one autoregressive lag and two lags of moving averages. An ARIMA model with robust standard errors yields an insignificant negative effect ($P = 0.219$) of the first difference of the relative frequency of *Heldentum* on the first difference of the *POLITY2* score. However, fitting the same model with 10 lags of autocorrelations and 1 lag of moving averages yields a significant negative effect ($P = 0.016$), but the fact that it requires many iterations to converge indicates that the model is misspecified.

In a similar vein, we can check the autocorrelations and partial-autocorrelations and then fit an ARIMA of the first difference of *Zürich* on the first difference of the *POLITY2* score and include two lags of autocorrelations and one lag of moving averages. This yields an insignificant negative effect ($P = 0.566$) of the first difference of the POLITY2 score on the relative frequency of *Zürich*. If we fit the model again and include nine lags of autocorrelations and two lags of moving averages, then we obtain a significant negative effect ($P = 0.032$), again with many iterations to converge.

These differences demonstrate why it is very difficult to choose the 'best' model specification in time-series analysis. That is why Becketti (2013, p. 268, my emphasis) issues a warning: '[t]ime-series analysis provides powerful tools for revealing patterns and relationships in data, but the best statistical techniques can only bound, but not eliminate, the irreducible uncertainty we face when analyzing data. [...] There is no substitute for a thoughtful approach to time-series analysis informed by deep subject-matter knowledge and willingness to apply rigorous tests to *every estimate*'. I believe that the analyses of Caruana-Galizia (2015) would certainly benefit from the identification of an appropriate ARIMA structure for 'every' keyword.