

# **Analyzing lexical change in diachronic corpora**

**Inauguraldissertation zur Erlangung des akademischen  
Grades eines Doktors der Philosophie der Universität**

**Mannheim**

vorgelegt von

**Alexander Koplenig**

am 17.02.2016

Datum der Disputation: 11.05.2016

Dekan: Prof. Dr. Matthias Kohring

Erstgutachter: Prof. Dr. Stefan Engelberg

Zweitgutachter: Prof. Dr. Martin Hilpert

## Table of contents

<b>Table of contents</b> .....	<b>1</b>
<b>List of figures</b> .....	<b>3</b>
<b>List of tables</b> .....	<b>4</b>
<b>Organization of the thesis</b> .....	<b>5</b>
<b>Acknowledgments</b> .....	<b>7</b>
<b>1 Introduction</b> .....	<b>9</b>
1.1 Motivation .....	9
1.2 The theoretical ideal of statistical inference .....	10
1.3 Corpora as representative language samples? .....	17
1.3.1 <i>Internalized</i> language as an unobservable cognitive phenomenon can be represented by <i>externalized</i> language as an observable phenomenon.....	18
1.3.2 Language data is non-random – statistical approaches that assume randomness cannot be applied.....	23
1.3.3 Language as a whole can be theoretically defined as a gigantic library that contains all utterances produced by the speakers of a language .....	25
1.3.4 Random sampling has to be replaced by the idea of balancing.....	30
1.3.5 Statistical inference cannot be used to extend the quantities found in one corpus to the language it seeks to represent .....	36
<b>2 Analyzing (lexical) change in diachronic corpora</b> .....	<b>40</b>
2.1 Data.....	42
2.2 Measuring similarity in synchrony .....	44
2.3 Case study: Visualizing the difference between the two spoken parts of the BNC.....	50
2.4 Measuring corpus similarity in diachrony .....	53
2.5 Case study: Lexical changes in the COHA .....	54
2.6 Further applications .....	57
2.7 Summary of the studies .....	58
2.8 References .....	61
2.9 Appendix .....	70

<b>3</b>	<b>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. ....</b>	<b>77</b>
<b>4</b>	<b>Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions.....</b>	<b>78</b>
<b>5</b>	<b>Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis .....</b>	<b>79</b>
	<b>Eidesstattliche Versicherung .....</b>	<b>80</b>

## List of figures

Figure 1: Histogram of percentage differences in the simulation of 1,000,000 samples.....	14
Figure 2: Estimated required sample size as a function of the percentage difference between two groups ( $p < .01$ ).....	24
Figure 3: Result of the simulation - obtained p-values as a function of the number of sentences sampled from the spoken section of the library .....	34
Figure 4: Abstract visualization of the method adapted in this section.....	47
Figure 5: Measuring the distance between corpora with varying degrees of similarity (darker shades of gray indicate more similar corpora). Solid black lines: linear fit between the observed values (hollow circles). .....	50
Figure 6: Visualization of the differences between the two spoken-parts of the BNC. The size of the words is proportional to the (log of the) partial contribution of the respective word type. Lighter shades of gray indicate that the word type is more frequent in the demog sample, while black word types are more frequent in the <i>cg</i> sample. ....	52
Figure 7: Visualizing changing genre compositions of the COHA using the $V$ -method.....	55
Figure 8: Visualization of the differences between the 1810s and the 1820 in the COHA. The size of the words is proportional to the (log of the) partial contribution of the respective word type. Lighter shades of gray indicate that the word type is more frequent in the 1820s, while black colored word types are more frequent in the 1810s. The labels on the y-axes of the small multiples show maximum and minimum values in the 1 million token samples (cf. Section 2.1). ....	56
Figure 9: Visualization of the differences between the time spans 1955-1959 and 2000-2004 in the American English GBC .....	70
Figure 10: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the British English GBC .....	71
Figure 11: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the French GBC .....	72
Figure 12: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the German GBC .....	73
Figure 13: Visualization of the differences between the time spans 1955-1959 and 2000-2004 in the Italian GBC .....	74
Figure 14: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the Spanish GBC .....	75
Figure 15: Visualization of the differences between the time spans 1950-1954 and 2000-2004 in the British English GBC .....	76

## List of tables

Table 1: Hypothetical relationship between color and the number of petals.....	11
Table 2: True relationship between color and the number of petals.....	13
Table 3: Result of a fictional study.....	33

## Organization of the thesis

This thesis consists of the following three papers that all have been published in international peer-reviewed journals:

- Chapter 3: Koplenig, Alexander (2015c). The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII. Published in: *Digital Scholarship in the Humanities*. Oxford: Oxford University Press. [doi:10.1093/llc/fqv037]
- Chapter 4: Koplenig, Alexander (2015b). Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. Published in: *Digital Scholarship in the Humanities*. Oxford: Oxford University Press. [doi:10.1093/llc/fqv030]
- Chapter 5: Koplenig, Alexander (2015a). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis. Published in: *Corpus Linguistics and Linguistic Theory*. Berlin/Boston: de Gruyter. [doi:10.1515/cllt-2014-0049]

These three chapters, as well as Chapter 2, which contains material taken from an earlier version of a paper and that is currently being reviewed by the *Journal of Quantitative Linguistics* (Koplenig, under review), form the main part of this dissertation. Chapter 1 introduces the topic by describing and discussing several basic concepts relevant to the statistical analysis of corpus linguistic data. Chapter 2 presents a method to analyze diachronic corpus data and a summary of the three publications. Chapters 3 to 5 each rep-

resent one of the three publications. All papers are printed in this thesis with the permission of the publishers.

## Acknowledgments

As Beckett puts it (2013: xxvii):

"Just once, I would like to see an author say, 'I did it all on my own, with no help from anyone else'. Of course, it would not be true, but it would be fun."

It goes without saying that this work would also not have been possible without the support of several people. I would first like to thank my advisor (and boss) Stefan Engelberg for his support, encouragement and persistent confidence in my ideas. The same can be said of my project leader, Carolin Müller-Spitzer, who from the very beginning supported my research in every possible way. I would also like to acknowledge and thank Martin Hilpert, who agreed to supervise my thesis on our first encounter. I am deeply indebted to my employer, the Institute for the German language (IDS) for generously providing me with all the hard- and software equipment I needed to complete this thesis. In particular, I would like to thank the director of the IDS, Ludwig M. Eichinger for his support well beyond the call of duty. I also wish to express my gratitude to my colleagues Marc Kupietz, Peter Meyer, Frank Michaelis and Sascha Wolfer for all the helpful discussions and insightful comments regarding the topics presented in this thesis. I am also grateful to Sarah Signer and Maria Parasca for proofreading and supporting this thesis.

*Darüber hinaus danke ich meinen Freunden und meiner Familie, insbesondere meinen beiden Eltern Marlyn und Wilhelm, ohne die all das nicht möglich gewesen wäre.*

Îi dedic această lucrare partenerei mele:

Andreea, îți mulțumesc din suflet că mi-ai fost și de această dată alături.

# 1 Introduction

Cuando se proclamó que la Biblioteca abarcaba todos los libros, la primera impresión fue de extravagante felicidad. [...]

A la desaforada esperanza, sucedió, como es natural, una depresión excesiva.

Jorge Luis Borges – La Biblioteca de Babel (1941)

## 1.1 Motivation

The idea for this thesis first emerged when the Culturomics team, in collaboration with Google, made its huge Google Ngram diachronic corpora (GBC) available for public use in 2010 (Michel et al. 2010). At that time, I hoped that this vast amount of data would enable me to study linguistic and cultural change with unprecedented accuracy as the 2009 version contains roughly 4% of all books ever published while the 2012 version includes a staggering 6% (Lin et al. 2012). The initial plan was simple: find traces of linguistic change by using state-of-the-art quantitative statistical methods. However, as it turned out, the plan was *too* simple, which was mainly the result of my poor knowledge of corpus linguistics at that time (hopefully), but, in my opinion, also has to do with the fact that most quantitative statistical methods used to analyze corpus data were initially developed in scientific disciplines other than (corpus) linguistics. To cut a long story short, the assumptions that allow inferences about a given population – in this case about the studied languages – based on results observed in a sample – in this case a collection of naturally occurring language data – are not fulfilled.

In this chapter, I therefore want to initially describe several basic concepts that are relevant for the statistical analysis of corpus data (Section 1.2). I will then show why the underlying methodological assumptions are not fulfilled in corpus linguistics by surveying several propositions that can be found in literature on the topic of representativeness (Section 1.3). On that basis, I will try to occupy a mediating position between (i) demonstrating why it is important to accept the fact that diachronic corpora are not representative in any sense and (ii) arguing that they still constitute a valuable sample of the written language record that can be used to understand the dynamics of linguistic change. To this end, the mathematical background of a method to measure (dis-)similarity in synchrony is adapted and extended to measure similarity in diachrony in Chapter 2. Section 2.7 consists of a summary of the three publications that together form the main contribution to this dissertation. As mentioned above, the three publications can be found in chapters 3 to 5.

## **1.2 The theoretical ideal of statistical inference**

Many empirical research projects face the problem that it is not possible or far too expensive to study the whole population i.e. all objects of interest, e.g. all citizens of a country, all animals of a given species or all stars in the Milky Way. Fortunately, it is not necessary to investigate all items of the population in the majority of situations. In this context, the main idea behind statistical frequentist inference is to use the distributional information from a sample of objects in order to estimate the characteristics of the unknown population from where the sample was taken. This is possible because under certain circumstances (discussed below) probability theory can be used to show that the distribution function of the population can be approximated by the distribution function

of the sample (Jann 2005: 124–127). The theory behind this rests on the assumption that the elements of the sample are chosen randomly from the population<sup>1</sup>:

"Conventional statistical inferences (e.g., formulas for the standard error of the mean, *t*-tests, etc.) depend on the assumption of random sampling. This is not a matter of debate or opinion; it is a matter of mathematical necessity." (Berk and Freedman 2003: 2)

As an example, we could consider a study of a certain type of flower in a certain rain-forest. We might be interested in color distribution (red or blue) and the number of petals (four or five). To this end, we could randomly pick 100 of these flowers in this rain-forest and find out that flowers with four petals are less likely to be blue (roughly 29 out of 100), than flowers with five petals, which are either red or blue (even split). Table 1 summarizes this result. The percentage difference is (50.0 % - 28.6 % =) 21.4 percentage points.

		Number of petals		
		four	five	
Color	blue	20 (28.6 %)	15 (50.0 %)	35 (35.0 %)
	red	50 (71.4 %)	15 (50.0 %)	65 (65.0 %)
		70 (100.0 %)	30 (100.0 %)	100 (100.0 %)

**Table 1: Hypothetical relationship between color and the number of petals**

<sup>1</sup> In this context, probability sampling refers to a situation in which the probability of any unit drawn into the sample is computable. The inverse probability can be used to weigh each sampled unit in order to approximate a representative sample (Berk and Freedman 2003: 15).

As in most cases of sample-based empirical research, we are actually less interested in the specific sample but instead want to make generalizations about e.g. the population of this type of flower in the rainforest. In our hypothetical case this would be that in this habitat, flowers with five petals tend to be red more often than flowers with four petals. But since we randomly sampled flowers, how can we be sure that the relationship observed in the sample also holds true for the entire population? The answer is, of course, that we can never be certain<sup>2</sup>; that the observed relationship is just the result of a biased sample, because we accidentally collected a disproportionate number of red flowers with four petals or a disproportionate number of blue flowers with five petals, or even both. However, statistical inference can help us judge this situation by quantifying the probability of a biased sample. In terms of statistical theory, we can calculate the probability of observing a relationship in a sample even though the relationship does not exist in the population of interest. Let us illustrate this notion with the help of a thought experiment: First, we assume that we already know that there is no relationship between the number of petals and the color of the flower in the populations because we went to the trouble of gathering the information for all, say, 1,000,000 flowers of the particular species in the rainforest.

Apart from random fluctuations, most flowers tend to be red, no matter how many petals the flower has. Table 2 summarizes this result.

---

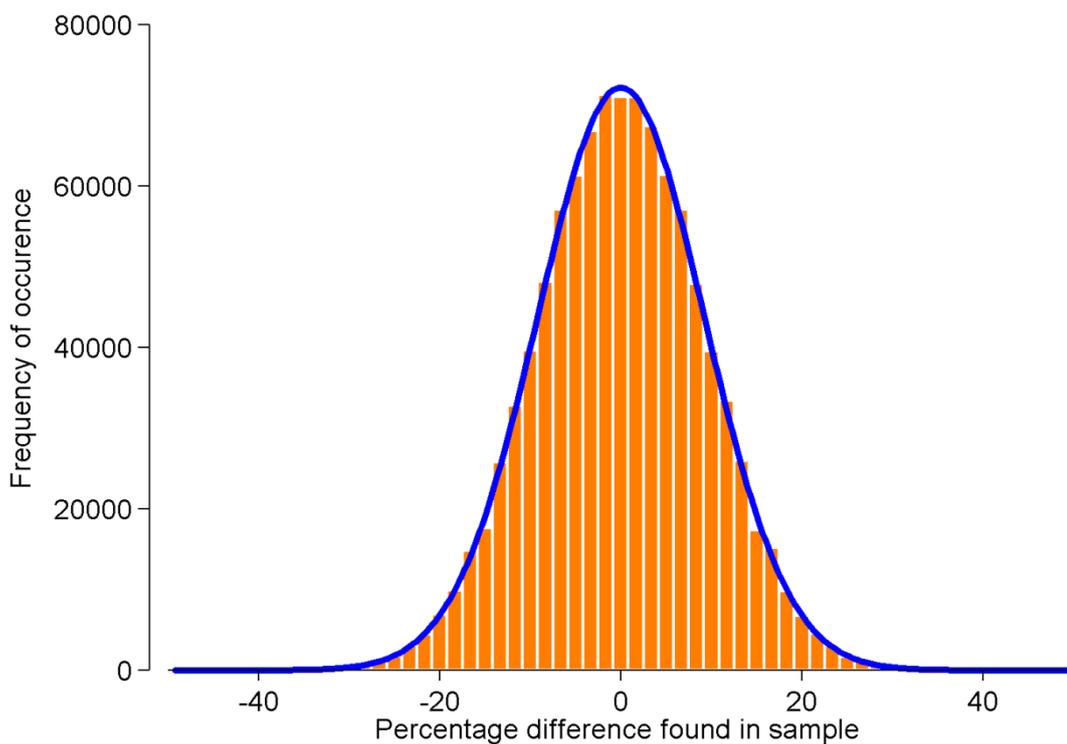
<sup>2</sup> It may be worth pointing out that this is not completely true. Based on our test sample, we can, of course, provide some certain information about the population: for example, we can rule out the hypothesis that all red colored flowers have four petals and so on. This is Popper's solution to Hume's problem of induction: "the assumption of the truth of test statements sometimes allows us to justify the claim that an explanatory universal theory is false" (Popper 1972: 7).

	Number of petals		
	four	five	
blue	150,000 (30 %)	150,000 (30 %)	300,000 (30 %)
red	350,000 (70 %)	350,000 (70 %)	700,000 (70 %)
	500,000 (100 %)	500,000 (100 %)	1,000,000 (100 %)

**Table 2: True relationship between color and the number of petals**

We could then repeat the data collection step plenty of times resulting in numerous separate samples, e.g. 1,000,000 different samples of 100 flowers. In each case, we could calculate the relationship of interest between color and the number of petals as in Table 1. Most of the time, we would not find an association between both variables which – as we already know – is the correct result. However, there would be a few cases where we might find a relationship similar to that described in Table 1. The idea of statistical significance follows from this argument: A result found in a sample is considered statistically significant if the probability of observing such an effect (given that it does not actually exist in the population of interest) is smaller than or equal to a chosen level of significance, for example 5%. In our example, this means that the number of samples in which we find an apparent relationship must not exceed 50,000 of all 1,000,000 samples. If we choose a lower level of significance, e.g. 1%, then we would only accept results that are observable in 10,000 samples. The *p-value* in this context is the probability of observing a result that is equal, or even more extreme, than the one we found in our sample, given the fact that there is actually no relationship in the population. A Fisher's exact test for significance yields a  $p = 0.066$  for Table 1. This means that, if we

repeat the sampling numerous times, then the number of samples from which we obtain a result is similar (or more extreme) to that in Table 1 - even so, Table 2 reveals that there is no relationship in the population, should not exceed 66,000 out of 1,000,000 cases. To demonstrate the utility of this idea, I sampled 100 "flowers" from the population of 1,000,000 flowers (cf. Table 2); this step was repeated 1,000,000 times, resulting in 1,000,000 different sample compositions. Figure 1 depicts the result: In 19,564 of all 1,000,000 samples, the difference in percentage is equal to or bigger than that found in Table 1. As indicated by the test for significance, the number of biased samples is therefore lower than 6.6 %. In roughly 72% of all samples, the percentage difference is smaller than 10 percentage points.



**Figure 1: Histogram of percentage differences in the simulation of 1,000,000 samples.**

In this context, the so called "null hypothesis" states that there is no relationship between the measured quantities in the population, while its "rival", the "alternative hypothesis" assumes that there is a relationship. A result based on a sample is then called statistically significant if the probability of rejecting a true null hypothesis in favor of the alternative hypothesis is lower than or equal to a pre-selected threshold, the level of significance.

In general, the probability of finding results that falsely imply a relationship depend on:

- (i) the magnitude of the observed difference (also called effect size) and
- (ii) the sample size.

(i) can be explained with the help of the following example: If all blue flowers in our sample of 100 specimen have five petals and all red ones have four petals this might still be the result of a biased sample, but it is highly unlikely. Not a single sample of all 1,000,000 samples in the simulation led to this result. However, it also demonstrates that we can be incredibly unlucky: there is one extremely biased sample in which 27 of the 44 flowers with four petals are blue (61.36 %), while only 9 of the 56 flowers with five petals are blue (16.07%). This implies an obvious relationship between the number of petals and the color of the flower, although this relationship, as we already know, does not exist in the population.

(ii) makes sense because, if we were to increase the size of our sample and gather the information for 10,000 instead of 100 flowers finding false results would also become more unlikely. The rationale behind this idea can be illustrated with the help of two extreme cases:

- (1) We sample four flowers and find the following effect: two red flowers have five petals; two blue flowers have four petals. However, the probability of find-

ing such an "effect", despite the fact that it does not exist in the population, is very high: there are many differently composed samples where false positive results would be found.

(2) We sample all but one specimen of the 1,000,000 flowers in the rainforest. In this situation, it is intuitively plausible to accept this result as statistically significant no matter how small the actual effect is, because it is overwhelmingly unlikely or – in this case – impossible to find a hypothetical sample that would not show this effect.

However, (2) also implies that with increasing sample sizes, arbitrary small effects will be found to be statistically significant.<sup>3</sup> Some consequences of this fact that affect the analysis of corpus data will be described in Section 1.3. Hitherto, it is necessary to briefly outline some of the basic concepts of corpus linguistics in the next section.

To sum up this section, probability theory provides a solid theoretical basis for the process of estimating unknown population quantities based on the characteristics of a sample from it. As previously explained, this process rests on the assumption that the selection is carried out randomly. In the next section it will be argued that this key assumption is not given for language samples. As I will try to show, this is both a matter of principle and also has to do with the fact that the statistical understanding of representativeness is widely rejected in corpus linguistics and replaced with the idea of balancing, i.e. including a large variety of different texts. However, if the traditional notion of representativeness is rejected in corpus linguistics, then everything that is based on this notion – especially basic significance testing – has to be rejected, too. A corpus sample is not representative – in a statistical sense – of the population and no statistical method can compensate for this problem.

---

<sup>3</sup> It is worth noting that testing for statistical significance becomes superfluous when all members of a population are surveyed.

### 1.3 Corpora as representative language samples?

Corpus linguistics can be thought of as the discipline that studies language "based on examples of 'real life' language use" (McEnery and Wilson 1996: 1). On that basis, a (synchronic) corpus can be defined as:

"a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety" (McEnery et al. 2006: 5; for a similar definition see Gilquin and Gries 2009: 6).

Condition (2) is based in the idea of analyzing "naturally occurring data" (Gries 2006: 4), while condition (1) refers to the advantages of modern computer technology (McEnery and Wilson 1996: 23–24): for example, machine-readable text collections can be quickly searched and manipulated. Furthermore, the corpora can easily be enriched with additional information, e.g. part-of-speech tags (POS). Given the vast size of today's corpora, it is not possible to manually inspect the corpus however; Michel et al. (2010: 176) summarize the following for the GBC that form the main data basis of this dissertation:

"If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years."

In the following, I want to discuss condition (3) and (4) by surveying several propositions that can be found in literature on the topic of representativeness:

- (i) *Internalized* language as an unobservable cognitive phenomenon can be represented by *externalized* language as an observable phenomenon, that is, real-life usage attested in corpus data (cf. Section 1.3.1).

- (ii) Language data is non-random; therefore statistical approaches that assume randomness cannot be applied (cf. Section 1.3.2).
- (iii) Language as a whole can be theoretically defined as a gigantic library that contains all utterances produced by the speakers of a language; corpora can then be considered representative samples drawn from this library (cf. Section 1.3.3).
- (iv) Random sampling has to be replaced by the idea of balancing, i.e. including a large variety of different texts in order to represent a language as a whole (cf. Section 1.3.4).
- (v) Corpora are not random samples of a language; therefore statistical inference cannot be used to extend the quantities found in one corpus to the language it seeks to represent (cf. Section 1.3.5).

### **1.3.1 *Internalized* language as an unobservable cognitive phenomenon can be represented by *externalized* language as an observable phenomenon**

In most situations, linguists are not interested in the specific texts included in a corpus, but instead want to find generalizations about the studied language and its structure (Baroni and Evert 2009; Evert 2006; Kohonen 2007; Leech 1991). In this context, a dichotomy famously put forward by Chomsky (1986) seems especially relevant: Chomsky distinguishes between *internalized* language (*I-language*) and *externalized* language (*E-language*). *I-language* can be considered "some element of the mind of the person who knows the language, acquired by the learner, and used by the speaker-hearer." (Chomsky 1986: 22), while *E-language* defines "a language as a collection of actions, or utterances, or linguistic form (words, sentences) paired with meaning, or as a system of linguistic forms or events" (Chomsky 1986: 19). One of the most basic ideas of corpus linguistics is that *I-language*, as a cognitive and therefore unobservable phenomenon that focusses on "the properties of a language as a formal system" (Evert 2006: 178)

can be represented by *E-language* "defined as the set of all utterances produced by speakers of the language" (Baroni and Evert 2009: 1) and is therefore a measurable and observable phenomenon. However, this "performance-based orientation towards language" (Leech 2007: 3) is strongly criticized by Chomsky, who believes that "E-language that was the object of study in most of traditional or structuralist grammar or behavioral psychology is now regarded as an epiphenomenon at best" (Chomsky 1986: 25). Chomsky (1986: 26–27) proposes a

"conceptual shift from E-language to I-language, from behavior and its products to the system of knowledge that enter into behavior [...W]hen we speak of a person as knowing a language, we do not mean that he or she knows an infinite set of sentences, or sound-meaning pairs taken in extension, or a set of acts or behaviors; rather, what we mean is that the person knows what makes sound and meaning relate to one another in a specific way, what makes them 'hang together', a particular characterization of a function perhaps."

In his opinion, the shift in focus from *E-language* to *I-language*, "from the study of language regarded as an externalized object to the study of the system of knowledge of language attained and internally represented in the mind/brain" (Chomsky 1986: 24) is a logical step, because languages in the *E* sense "are not real-world objects but are artificial, somewhat arbitrary, and perhaps not very interesting constructs" (Chomsky 1986: 26).<sup>4</sup>

---

<sup>4</sup> There is a certain irony in the fact that regarding this cognitive turn, Chomsky stated elsewhere that "[i]t is quite possible – overwhelmingly probable, one might guess – that we will always learn more about human life and human personality from novels than from scientific psychology" (Chomsky 1988: 159), because this is exactly what corpus linguists hope for when they use natural language data (including novels) to gain insights about the language faculty.

Before assessing Chomsky's view and putting it into perspective of contemporary corpus linguistics, it seems fair to point out that in the 1950s, when Chomsky first argued that quantitative data is of no use to linguistics (McEnery and Wilson 1996: 4–11), his criticism was aimed at the corpora of the time, which predominantly consisted of "shoe-boxes filled with paper slips" (McEnery et al. 2006: 3). With the advent of powerful computers with which to store, process and analyze incredibly large amounts of textual data, however, Chomsky's criticism has to be re-evaluated: the claim that there is no connection between *I-* and *E-language*, or put differently, that *E-language* cannot be used to learn anything *at all* about *I-language* is quite a counterintuitive and strong claim. And as the famous cosmologist Carl Sagan once said: "Extraordinary claims require extraordinary evidence". So, I believe that Leech (2007: 3) is right to assume that "E-language is a crucial, indispensable manifestation of I-language". Someone who doubts this would have to collect *extraordinary* evidence to support his or her doubts. Chomsky himself does not present this evidence but simply advocates the usage of the intuition of one native speaker (himself in this case) as the "empirical" basis for the study of *I-language*:

"In actual practice, linguistics as a discipline is characterized by attention to certain kinds of evidence that are, for the moment readily accessible and informative: largely, the judgments of native speakers. Each such judgment is, in fact, the result of an experiment, one that is poorly designed but rich in the evidence it provides.

[...T]he judgments of native speakers will always provide relevant evidence for the study of language, just as perceptual judgments will always provide relevant evidence for the study of human vision." (Chomsky 1986: 36–37)

Wasow and Arnold (2005) argue that it is unproblematic to use intuitions "as evidence for theoretical claims", for example when introspectively judging how well-formed a given expression is. However, they also demonstrate that "intuitions about why a given expression is (or is not) well-formed or has the meaning it has [...] do not themselves constitute evidence for or against theoretical claims", but are only one source of evidence. In addition, Schütze (1996) showed that such introspective judgments are by no means unbiased, objective or reliable as assumed by Chomsky and his followers. Therefore, intuitions "should have no privileged status relative to other forms of evidence" (Wasow and Arnold 2005: 1485; see also Gilquin and Gries 2009).

Nonetheless, I believe Váradi (2001: 587) is right to claim that Chomsky's dichotomy is one that corpus linguistics "has to face", because *I-language* is a cognitive phenomenon and as such is not directly observable. Angrist and Pischke (2008: 24) point out that we "must [first] define the objects of interest before we can use data to study them". Therefore, it is simply not sufficient to stipulate that approximating *I-language* by *E-language* is possible "with all the paradoxes that this view implies" (Baroni and Evert 2009: 1) or merely note that "statistical inference, on the other hand, will not be of help in solving thorny issues such as what is the appropriate extensional definition of a 'language as a whole' and how we can sample from that" (Baroni and Evert 2009: 1). For a better understanding of how corpus linguistic evidence can be interpreted in cognitive terms, we first need a better understanding of the relationship between corpus linguistic evidence and other sources of linguistic evidence, such as (psycholinguistic) experimentation (e.g. lexical decision tasks, eye-tracking studies), elicitation (sentence completion, sentence sorting, acceptability judgments) or neurolinguistic experimentation (Arppe and Järvikivi 2007a; Gilquin and Gries 2009: 5) in order to assess "the cognitive reality of

corpora" (Gilquin in Arppe et al. 2010: 6) and to "strengthen the empirical foundations of corpus linguistics" (Leech 2007: 134). Interestingly, even Chomsky agrees in this context:

"In principle, evidence concerning the character of the I-language and initial state could come from many different sources apart from judgment concerning the form and meaning of expressions: perceptual experiments, the study of acquisitions and deficit or of partially invented languages such as creoles, [...] or of literary usage or linguistic change, neurology, biochemistry, and so on." (Chomsky 1986: 36–37)

Promising case studies that combine corpus data with other empirical information can be found in Arppe & Järvikivi (2007a; 2007b), Baayen (2010), Gilquin (2008), Gries et al. (2005), Gries et al. (2010), Kertész & Rákosi (2008), Mandra et al. (2015), Schmid (2010), or Wiechmann (2008). On a more general level, such a multi-methodological and multi-disciplinary, and therefore costly, research agenda could not only be highly beneficial for corpus linguistics, but for linguistics in general. This has to do with the fact that language is a multimodal phenomenon:

"Languages are spoken and listened to, signed and watched, written and read, encrypted and decoded, studied and described, taught and learned, analyzed and generalized and imitated. Languages are represented as sounds or visual signs, scratched in a myriad of ways on a plethora of media. Languages are produced by an interaction of the lungs, vocal cords, throat, tongue, and mouth, or by the coordination of the hand(s). Languages are understood via the ears or the eyes, or even the skin, and in the end comprehended in the mind. Languages are uttered spontaneously for the instantaneous need of the moment, or recorded intentionally for the legacy of eternity. It should be obvious from this kaleidoscope of different representations and characteristics of human language that it is a multimodal phenomenon which we can expect to understand fully and

comprehensively only by combining multiple methods and multiple sources of evidence, by scientists and practices from multiple disciplines." (Arppe and Järvikivi 2007a; see also Arppe et al. 2010)

### **1.3.2 Language data is non-random – statistical approaches that assume randomness cannot be applied**

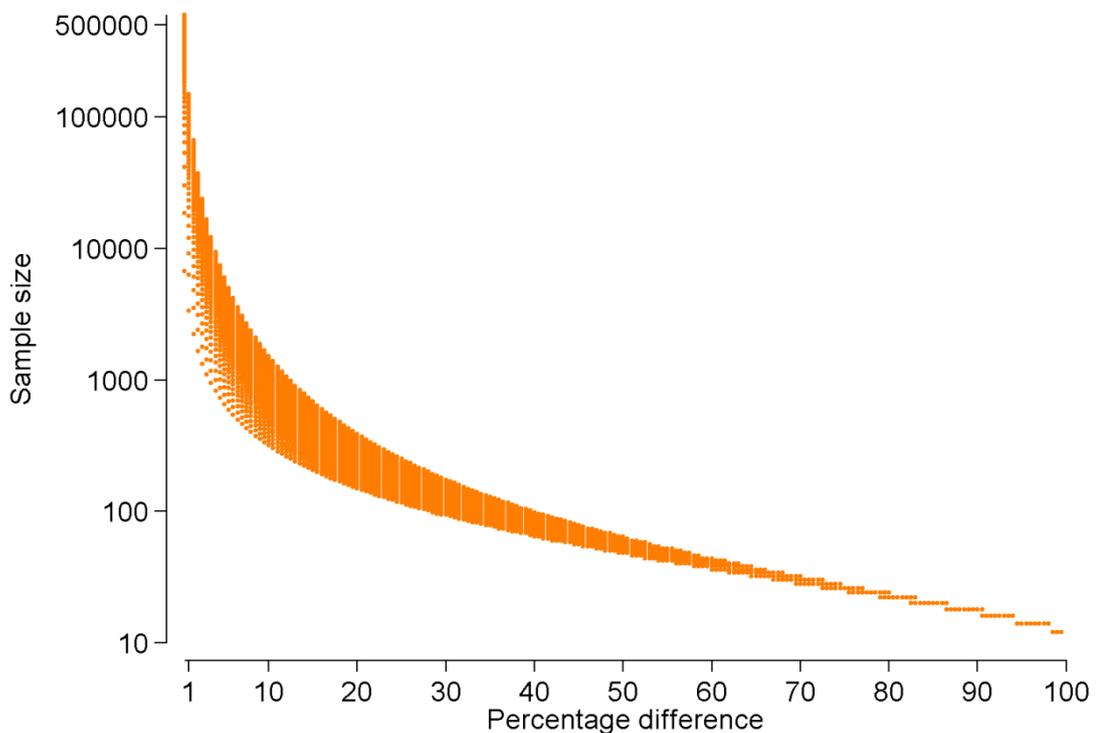
As explained in Section 1.2, a result is called statistically significant if the probability of rejecting a true null hypothesis in favor of the alternative hypothesis is lower or equal than the pre-selected level of significance. In an influential paper, Kilgarriff (2005: 273) argued that:

"Language users never choose words randomly, and language is essentially non-random. Statistical hypothesis testing uses a null hypothesis, which posits randomness. Hence, when we look at linguistic phenomena in corpora, the null hypothesis will never be true."

In my opinion, this argument has to be questioned because while it is certainly the case that words are not chosen at random, this does not affect the validity of a statistical test as it only assumes that the elements of a sample are randomly selected from the population. In reference to Evert (2006), the idea of a corpus as a random sample of 'a language as a whole' will be examined in more detail and critically discussed in the next section (1.3.3). A much more important point in Kilgarriff's paper is the observation that apart from the magnitude of the found effect, statistical significance depends on the size of the sample as explained above. With ever-growing available language data, this observation is of special importance for corpus linguistics, as Kilgarriff (2005: 263) puts it:

"In corpus studies, we frequently do have enough data, so the fact that a relation between two phenomena is demonstrably non-random, does not support the inference that it is not arbitrary."

To visualize this, Figure 2 plots the estimated required sample size in order to achieve statistical significance (at  $p < .01$ ) as a function of the percentage difference between two groups, for example two (sub)corpora. For means of simplification, it is assumed that both groups are of an equal size. The percentage difference and the estimated sample size is calculated for all possible combinations of different distributions of two groups (each ranging from 0 to 100 %; incremented by 0.5%).



**Figure 2: Estimated required sample size as a function of the percentage difference between two groups ( $p < .01$ )**

Figure 2 demonstrates that large percentage differences are needed in order to obtain statistical significant effects in the case of small samples. At the same time, it also clear-

ly shows that arbitrary differences become highly significant with increasing sample size, for example a difference of 50.05 % in the first group and 50.00 % in the second is found to be highly statistical significant for a sample size of roughly 600,000.

Consequently, in a follow-up on Kilgarriff, Gries (2005) proposes also taking the magnitude of an obtained effect into consideration by using measures that evaluate the relative size of an effect in relation to the available data. He then shows that using those measures can help "do away many null-hypothesis testing problems" and asks:

"Do the points of critique and the proposals [...] as well as the present findings also mean that we as corpus linguists should more or less abandon null-hypothesis significance testing?" (Gries 2005: 284)

I would go so far and say that: yes, significance testing should be abandoned in corpus linguistics. In the next three sections, I want to substantiate this claim.

### **1.3.3 Language as a whole can be theoretically defined as a gigantic library that contains all utterances produced by the speakers of a language**

In order to define *I-language* in terms of *E-language*, Evert (2006) argues that a corpus can, as mentioned in the last section, be considered a random sample of a language as a whole; he asks us to:

"imagine a gigantic library that represents the entirety of a language or sublanguage as the object of study. Each book in this library corresponds to a fragment of the language – some large, some small – that could be used as a linguistic corpus. Selecting or compiling a corpus, thus, amounts to picking a book at random from one of the shelves. In this way, randomness enters quantitative corpus studies, even if it is not inherent in the object of study itself, viz. the language under investigation" (Evert 2006: 178)

Since the unit of measurement, i.e. the quantity of interest (e.g. words, phrase or whole sentences), varies from case to case, Evert (2006) refines his thought experiment:

"Imagine that someone went through the library and cut every book into small paper slips, each one carrying a single token (word, phrase or sentence, depending on the unit of measurement). This would leave a big heap of paper slips, containing exactly the same words with exactly the same relative frequencies as the original library. Instead of picking a book from one of the shelves, we can now take a handful of paper slips from this heap, giving us a *random sample* of tokens from the library." (Evert 2006: 182)

I believe that this idea could indeed be used as a theoretical principle in order to define a corpus as a random sample and therefore treat it *as if* it were a representative sample of the language it seeks to represent (Berk and Freedman 2003). However, in this section I will try to show that the key assumptions behind this idea are never fulfilled in practice for reasons of principle.

A first problem is mentioned by Evert (2006: 185):

"sampling at the unit of measurement, i.e. individual words or sentences from the entire library [...] is impracticable because it would require each word or sentence to be taken from a different book. [...] Just imagine how difficult it would have been to compile [a] corpus by sampling one word each from a million books, rather than taking 2,000-word samples from only 500 books."

But just because something is impracticable (which it certainly is) does not imply that we can just gloss over this problem and assume that it does not *really* matter (which it certainly does, cf. Váradi 2001). Using more appropriate methods that do not assume independence at the text-level can be found in Brezina and Meyerhoff (2014), Gries (2015), or Lijffijt et al. (2014). It is important to keep in mind, however, that those

models rely on the assumption that the texts which the corpus compiles are random samples from the textual universe, that is the gigantic library approximating language as a whole (this in turn counters the idea of balancing a corpus, i.e. subjective text selection, cf. Section 1.3.4). It seems equally important to understand that Evert (2006) does not mean "book" in a conventional sense. Each "book" corresponds "to a fragment of the language". So in principle, a "book" could also be a transcript of spoken language. However, as a result of (i) legal restrictions, (ii) problems of data collection and (iii) obtrusiveness, many types of spoken data just cannot be collected. (i) refers to the fact that – for very good and obvious reasons – in many countries, language data cannot be recorded without the "informed consent" of the speakers (Deppermann and Hartung 2011). Assuming that all (!) speakers of a language (including presidents and mobsters) were to give us their permission to record all (!) of their spoken data (I definitely would not), and assuming that there would be no ethical problems (Deppermann and Hartung 2011: 443), we would still have to solve (ii). Since this point is only (!) a practical problem, and could in principle be solved by equipping each speaker of a language with a language data recorder and then automatically or manually transcribing each recorded fragment, we are left with (iii) which is the most principled problem in this context. Obtrusive measurement means that the researcher has "to intrude in the research context" (Trochim 2006). If people know that they are being recorded, this information is likely to influence their behavior (Kellehear 1993: 5), e.g. when talking about intimate preferences or planned crimes. This results in a situation where the recorded data lose their authenticity (Deppermann and Hartung 2011: 444).

Therefore, we would have to face the fact at this point that our gigantic library is biased towards written language. If we assume that this is not a problem (which it certainly is),

we could restrict our library to written language by stipulating that the entire written language record can be used to approximate language as a whole. However, this restriction results in some severe problems, too: if we assume that all publishers were to give us access to (and permission to use) their data, we would still have to admit that many types of written language fragments are not published, for example shopping lists and love or blackmail letters. So, after wiring all speakers of a language, we would also have to make sure that they keep everything they have ever written down. Even if we assume, for the sake of the argument, that we could accomplish this, we would be confronted with similar problems to the ones discussed above for spoken language because the outlined procedure would affect the authenticity of the texts: if a person knows that an intimate note to his or her partner is being stored, how can we make sure that she or he does not leave out certain details or – even worse – that she or he writes the note at all.

Another problem arises that is not quite as obvious: what do we actually mean by all fragments of a language, i.e. the set of all utterances produced by the speakers of a language (cf. Section 1.3.1)? For example, what should we do with in-text quotes, or abstracts, or subheadings that often repeat parts of the text? Or even more importantly – given the fact that many contemporary corpora predominately consist of newspaper texts – what should we do with stories and reports provided by (international) news agencies that are bought and published by several (regional) newspapers? Do they count as separate utterances or only as one utterance?

In addition: what happens with draft versions of a text, for instance a draft of a newspaper article that is "heavily edited by editors and type setters for reasons that [...] may or

may not be linguistically motivated" (Gilquin and Gries 2009: 7) – should we include the draft, or the final version, or both?

Furthermore, should the library consist of all different types or tokens of "books"? So should we include all printed copies (or tokens) of a bestselling book or only one type. The former seems to be preferable in this context, because a bestselling book like one of the Harry Potter series is likely to affect both language reception and production to a greater extent than a non-seller such as, say, this thesis (Leech 2007: 7). But what about e-books? And, is the number of printed copies really an unbiased indicator? Because one book or paper can be read by more than one person or several times by one person. And, just because a book is being sold, does not necessarily mean that it is also being read: in an essay that was published in the Wall Street Journal, Ellenberg (2014) tries to demonstrate that many bestselling e-books are "left unfinished". The same could be said about:

"a radio programme that is listened to by a million people should be given a much greater chance of being included in a representative corpus than a conversation between two people, with only one listener at any one time."  
(Leech 2007: 6)

Again, how can we make sure that all one million people *really* listened (the whole time)?

To be fair, Evert's library serves only as an illustration. So, instead of taking it at face value as I did in this section, maybe it would be better to think about language as a whole as an imaginary population. However, as Berk and Freedman (2003) argue, postulating an imaginary population from which the given data is assumed to be randomly drawn is essentially circular and makes it necessary to:

"demonstrate that the data can be treated as a random sample. It would be necessary to specify the social processes that are involved, how they work, and why they would produce the statistical equivalent of a random sample. **Handwaving is inadequate.** [...] The rhetoric of imaginary populations is seductive precisely because it seems to free the investigator from the necessity of understanding how data were generated. " (Berk and Freedman 2003: 4, my emphasis)

In my opinion, the points discussed in this section demonstrate that, in this context, Evert's (2006) library-metaphor is very good: not in order to show how the random-sample-model of statistical analysis can be applied in corpus linguistics, but to illustrate the problems that come with defining language as a whole *as if* it was an imaginary population where our corpus data is sampled from.

As a consequence, this could mean that corpus linguistics needs a different notion of representativeness. This leads me to the next section.

#### **1.3.4 Random sampling has to be replaced by the idea of balancing**

In a very famous and much cited paper on this topic, Biber (1993) tries to answer the of how a corpus can *represent* a language. He argues that:

"Representativeness refers to the extent to which a sample includes the full range of variability in a population." (Biber 1993: 243)

While this is not true from a statistical point of view, as shown in Section 1.2, it is important to emphasize that, as in any other scientific discipline, corpus linguistics naturally has the right to (re)define its key concepts in a suitable way in order to fulfill the special needs of the respective field of study. If, however, corpus linguistics requires a different "notion of representativeness" as Biber (1993) puts it, then it is completely unclear why that which is based on the traditional notion of representativeness (especially

the concept of statistical significance) can be used regardless of this "different notion".

Therefore, I completely share Váradi's negative view of Biber's approach:

"one must voice serious misgivings about any attempt to divest such a key term of its well-established meaning, which has a clear interpretation to statisticians and the general public alike. Of course, any self-respecting corpus would like to advertise itself as a representative corpus. There is such a strong and unanimous expectation from the public and scholars alike for corpora to be representative that it is an assumption that is virtually taken for granted. However, to meet this demand by the semantic exercise of redefining the content of the term is a move that hardly does credit to the field." (Váradi 2001: 592)

Biber's idea of representativeness refers to the idea that a corpora "include the full range of linguistic variation existing in a language" (Biber 1993: 247):

"Whether or not a sample is 'representative', however, depends first of all on the extent to which it is selected from the range of text types in the target population; an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from that population." (Biber 1993: 243)

In a similar vein, Perkuhn et al. (2012: 47) suggest that:

"In order to approximate representativeness or at least make it assessable, the distribution of dimensions in the corpus is normally controlled for; these dimensions are (i) **intuitively considered relevant** for the domain (or are expected to affect the outcome) and (ii) can be collected with an acceptable amount of effort." (my emphasis)<sup>5</sup>.

---

<sup>5</sup> My translation of "Um die Repräsentativität zu approximieren oder zumindest einschätzbar zu machen, wird bei der Korpuskomposition meist die Verteilung bezüglich solcher Dimensionen kontrolliert, die man (i) für die Sprachdomäne **intuitiv als relevant** erachtet (bzw. die voraussichtlich Auswirkungen auf Befunde haben werden) und (ii) mit vertretbarem Aufwand in Erfahrung bringen kann."

Again, there is nothing to be said against this strategy, but it also implies that the idea of statistical significance cannot be applied in corpus linguistics; intuition and statistical rigor are simply not compatible.

In the last section, based on Evert's library metaphor, it has been demonstrated that such an approximation or a "prior full definition" in terms of Biber is harder to reach than it might seem. To illustrate why I believe that balancing is rather problematic, take for instance one of the most famous corpora, the British National Corpus (BNC; Burnard 2007), a synchronic corpus consisting of 100 million words. Regarding the composition of the BNC, Evert (2006: 183; see also Leech 2007: 4) argues in reference to his library metaphor that:

"In a sense, a balanced corpus is representative of the relevant sub-language because it contains material from all the different sections of the library. However, one problem remains: in order to give an accurate picture of relative frequencies in the entire library, books must be selected in proportional numbers according to the relative sizes of the different sections. Without access to the full library, it is impossible to know the sizes of the sections, though. Hence, this step involves assumptions about how much material each section contributes to the library, i.e. assumptions that are necessarily subjective and often disputable. For instance, the BNC contains slightly more than 10% of spoken material. If BNC frequencies are taken to be representative of modern British English, there is an implicit assumption that only 10% of the output of British speakers consists of speech, while the remaining 90% are produced in writing. Based on this assumption, the frequency of passives in modern British English would be estimated to be 11.2 per 1,000 words (from relative frequencies of 12.1 in the written part and 4.2 in the spoken part of the BNC). It is quite likely that the true proportions are just the other way."

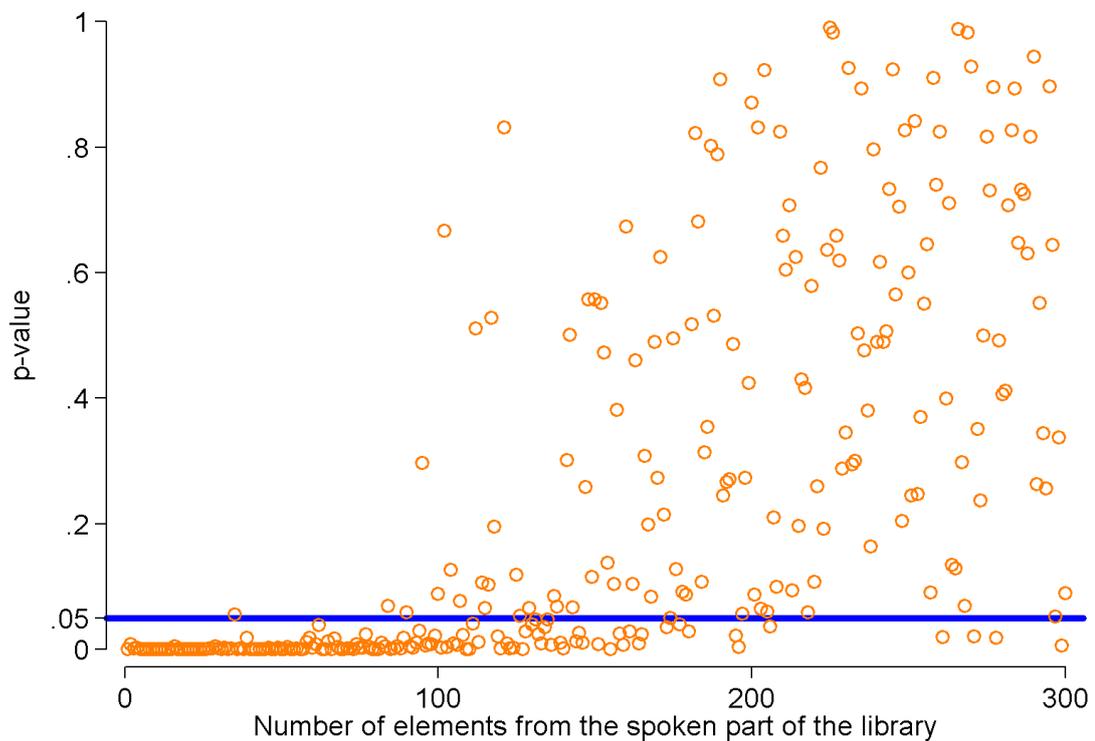
A small fictional example might be useful to illustrate this point. Let us assume that we are interested in a relatively rare phenomenon, let us further assume that the library were actually to exist and that, for pragmatic reasons (the distances in this library are vast), we would only be able to sample 300 units (sentences in this case). There are 580 sentences in the written section of the library and 758 sentences in its spoken section. In order to mimic the BNC sampling, we sample 90 % (370 sentences) from the written section of the library and 10 % (sentences) from the spoken section. Table 3 shows the result.

	X: present	X: absent	Totals
Y: present	47.86 %	28.13 %	37.33 %
Y: absent	52.14 %	71.88 %	62.67 %
Totals	100.00 %	100.00 %	100.00 %

**Table 3: Result of a fictional study.**

The result would lead us to believe that there is a strong relationship between the presence of X and the presence of Y in a sentence. When X is present, Y is present in 47.86 % of all instances. In the absence of X, Y is only present in 28.13 % of those cases. The difference of 19.73 percentage points is highly significant at  $p < 0.0005$  with a  $X^2$  of 12.43. To see what would happen if we changed our sampling scheme, I have written a short simulation in which the sampling is repeated 300 times. In the first case, we sample 0 sentences from the spoken section of the library and 300 sentences from the written section. In the second case, we sample 1 sentence from the spoken section and the

rest from the written section and so on. In the final case, we sample 300 sentences from the spoken section, and 0 sentences from the written section. In each case, we calculate the resulting  $p$ -value. Figure 3 presents the result. In roughly 53 % of all cases, the obtained  $p$ -value is larger than the common 0.05 threshold of significance (the blue line in the figure).



**Figure 3: Result of the simulation - obtained  $p$ -values as a function of the number of sentences sampled from the spoken section of the library**

At this point, statistical analysis does not help in determining which sampling distribution is most likely corresponds to the "true" distribution, i.e. the distribution in the popu-

lation.<sup>6</sup> From a statistical point of view, balancing is problematic because it is subjective per definition - different researchers might have different opinions on defining different registers and sub-registers as well as genres and subgenres that constitute different sections of our imaginary library. Thus, while Evert (2006: 184) claims that:

"Mathematically speaking, if each part of a balanced corpus is a random sample from the corresponding section of the library and if the relative sizes of these parts match the relative sizes of the library sections, then the whole corpus is a random sample from the entire library (at least to a very good approximation)",

I believe Váradi (2001: 590) to be right. Since we do not know the "true" proportions, we also do not know how to balance in a proportional and objective way. Without such a method, the statistical inferences from a balanced corpus to Evert's library that approximate language as a whole are invalid. On these grounds, the pessimistic view held by McEnery et al. (2006: 21) seems to be warranted:

"Claims of corpus representativeness and balance [...] should be interpreted in relative terms and considered as statement of faith rather than as fact, as presently there is no objective way to balance a corpus or to measure its representativeness" (cf. also Kohnen 2007)

This leads us directly to the next section.

---

<sup>6</sup> In this context, Evert (2006: 186) seems to assume the opposite: "The variation between samples taken from different sections of our metaphorical library is not a problem of the statistical analysis but of corpus design, which has to ensure that the composition of the corpus mirrors the sections of the library [...]. If the random sample model holds within each section, its inferences are also valid for a balanced corpus collected from the entire library." This comes as a bit of a surprise as he says elsewhere (as quoted in Section 1.3.1, in a paper co-authored with Marco Baroni) that: "statistical inference [...] will not be of help in solving thorny issues such as what is the appropriate extensional definition of a 'language as a whole' and how we can sample from that" (Baroni and Evert 2009: 1).

### **1.3.5 Statistical inference cannot be used to extend the quantities found in one corpus to the language it seeks to represent**

At first glance, it may seem trivial to assume that a corpus is a random representative sample of a particular language. A closer look outlined in the last four sections, however, revealed that there are many good reasons to doubt this assumption. What does this imply? Well:

"Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else."  
(Leech 2007: 135; see also Hunston 2010; Durrell 2015)

While this is arguably not a very pleasant consequence (Schönefeld 2011: 16; or Köhler 2005), it seems from a methodological point of view to be the only logical one:

"If the random-sampling assumptions do not apply, or the parameters are not clearly defined, or the inferences are to a population that is only vaguely defined, the calibration of uncertainty offered by contemporary statistical technique is in turn rather questionable." (Berk and Freedman 2003: 1)

It is worth emphasizing that this problem cannot be fixed by choosing statistical models that are better suited to the special properties of natural language data, especially models that do not rest on the assumption that the units of measurements (words, phrases, sentences, etc.) have to be independent within the texts included in the corpus (cf. Section 1.3.3). Those methods can and should be used to accurately describe the (linguistic) structure found in one corpus. However, because, as Baroni and Evert (2009: 2, see also Section 1.3.1) put it,

"[i]t is rarely the case that linguists are interested in the samples per se, rather than in generalizations from the samples to the infinite amount of text corresponding to the extensional definition of a (sub)language",

the inherent problem remains: a biased sample is a biased sample (and in the era of big data: this is true regardless of the size of the sample, cf. Koplenig and Müller-Spitzer, under review). No matter how you approach it, it seems that Rieger's claim (1979) is right: speaking about corpora as representative samples is inappropriate. That is why Gries's question (2005: 284 cf. also Section 1.3.3) could be answered: corpus linguists should indeed abandon significance testing.<sup>7</sup>

Does this imply that a question such as: "[i]s word  $X$  more frequent in male conversation than in female conversation?" (Lijffijt et al. 2014: 2) cannot be answered based on any corpus?<sup>8</sup> I believe it does, especially if this question is framed in terms of statistical

---

<sup>7</sup> In the context of corpus linguistics, this is not entirely true. Using appropriate techniques, significance testing can, of course, be used. It is the goal of a study to make generalizations about a large corpus based on a smaller random sample drawn from it (Oakes 1998: 10). On the other hand, it is also worth pointing out that in many cases, testing for statistical significance is not even necessary. Take for instance the example of McEnery and Wilson (1996: 69–71). As an illustrative example, they use the Latin versions of the Gospel of Matthew and the Gospel of John. They try to find out how often the present tense form of the verb "to say" (*dicit*) and how often the perfect form (*dixit*) is used. They then try to show how to calculate the statistical significance of the relationship by using a  $X^2$  test. They conclude that "the difference which we found between Matthew and John is significant at  $p < 0.05$ , and we can therefore say with quite a high degree of certainty that this difference is a true reflection of variation in the two texts and is not due to chance." (McEnery and Wilson 1996: 71). However, what is the population that the samples try to represent? If we were interested in general differences between the fictional works of Matthew and John, we could draw a random sample of all sentences with the respective verb and then conduct a significance test. However, as far as I know, the only texts that are attributed to Matthew and John are in fact the two Gospels. Thus, an analysis of those two works actually amounts to an exhaustive data collection process that is a census, and in this context a test for statistical significance does not make any sense since there is no inference (cf. fn. 3). Therefore, McEnery and Wilson (1996) can safely conclude that in his Gospel, John uses the present tense more often than Matthew does in his version.

<sup>8</sup> It may be worth pointing out that using significance testing in order to find the most "meaningful" differences between different (sub)corpora is problematic, too: since the process of selecting the strongest associations in a given dataset is selective, evaluating the strength of the obtained associations is tricky

significance. Thus, it might not be possible to find "out whether a word occurs significantly more often in one text or corpus than in another" (Lijffijt et al. 2014: 1), but it is, of course, possible to find out whether a word occurs more often in one (sub)corpus compared to another.

In order to find out something (potentially more interesting) about the studied language itself, Berk and Freedman (2003: 16) recommend a:

"better focus on the questions that statistical inference is supposed to answer. If the object is to evaluate what would happen were the study repeated, **real replication is an excellent strategy**. [...] Empirical results from one study can be used to forecast what should be found in another study." (my emphasis)

Or put differently, if we find an interesting result in one (sub)corpus, we can use this information to make predictions about another (sub)corpus or other types of linguistic data (for an overview see Gilquin and Gries 2009).<sup>9</sup> Again, this idea points towards the importance of converging evidence (cf. Section 1.3.1): If a result holds true across different corpora and – even better – for different types of linguistic data, we can use this form of *converging* evidence to cautiously postulate a general relationship – maybe even for the language as a whole.<sup>10</sup>

---

because this "cherry-picking" approach makes it necessary to modify the way statistical significance is calculated in post model selection inference, as demonstrated by Taylor and Tibshirani (2015).

<sup>9</sup> It is also interesting to note that this implies that, instead of using a frequentist approach to inference (as outlined in Section 1.2), corpus linguistics might benefit from choosing a Bayesian framework which offers a principled method to use prior information in order to make predictions about new empirical evidence (Thompson 2014).

<sup>10</sup> One might object that in many (if not most) cases of experimental (e.g. psychological) research, the participating subjects are mostly (undergraduate) students that do not form a random sample of the population. While this is certainly true, the key to thinking about experiments lies in the random assignment of the subjects to the different experimental conditions, because this design structure guarantees internal

Throughout the remainder of this thesis, I will try to occupy a mediating position between (i) demonstrating why it is important to accept the fact that diachronic corpora are not representative in any sense and (ii) arguing that they still constitute a valuable sample of the written language record that can be used to understand the dynamics of linguistic change. Before summarizing the studies, a few words about diachronic corpora are in order, followed by a description of a method with which to analyze (lexical) change.

---

validity (Campbell and Stanley 1966; Pearl 2009). This can be best understood in terms of placebo-controlled medical trials, where the respondents in the control group receive a treatment without any active substance. The measured effect in this group is then compared with the respondents in the experimental group who received actual treatment (Shang et al. 2005). Randomly assigning the units of interest to the experimental conditions eliminates the selection bias, which is the potential influence of confounding variables on an outcome of interest. Balancing the "subject's characteristics across the different treatment groups" (Angrist and Pischke 2008: 14) ensures that the experimental condition and all possible (even unidentified) variables that could affect the outcome are uncorrelated. Through the manipulation of the independent variables, it can be inferred that, random fluctuations aside, the treatment is the cause of the outcome (Diekmann 2002: 297): if the effect in the treatment group differs significantly (either positively or negatively) from the effect in the control group, then the only logical explanation for this is a causal effect of treatment on outcome. Therefore, I believe that Angrist and Pischke (2008) are right to say that "[t]he most credible and influential research designs use random assignment." (Angrist and Pischke 2008: 9). This again shows why the idea of testing corpus evidence experimentally could be highly beneficial for corpus linguistics, psycholinguistics and linguistics in general (Gilquin and Gries 2009).

## 2 Analyzing (lexical) change in diachronic corpora

In addition to the definition of a (synchronic) corpus that was presented in the last section, a diachronic corpus can be considered a collection of "texts from the same language gathered from different time periods" (McEnery et al. 2006: 65). It is worth noting that most (if not all) synchronic corpora consist of "texts that vary along the parameter of time" as indicated by Hilpert and Gries (2009: 400), because they contain texts from different periods of time. The specific idea of compiling a diachronic corpus is to explicitly focus on "the added parameter of time that must be adequately represented" (Biber 1998: 251) in order to analyze linguistic change and language evolution. In this context, McEnery et al. (2006: 96) are right to point out that:

"Diachronic study is perhaps one of the few areas which can only be investigated using corpus data. This is because the intuitions of modern speakers have little to offer regarding the language used hundreds or even tens of years before."

It is equally important to emphasize that, compared to synchronic corpus design and the problems outlined in Section 1.3, the situation is even more complex when designing a diachronic corpus for broader multi-purpose research goals (Biber 1998: 251–253). As Labov (1994: 11) famously stated:

"Historical documents survive by chance, not by design, and the selection that is available is the product of an unpredictable series of historical accidents."

It does not seem far-fetched to assume that historical documents do not only "survive by chance", but that the rate of survival is higher for documents (especially for older ones)

which were written by wealthy individuals with a good reputation, due to the fact that these individuals were both capable of writing documents and preserving them. This bias imposes an additional constraint to the analysis of diachronic data.

The recent availability of large machine-readable diachronic corpora such as the Corpus of Historical American English (COHA; Davies 2010a) or the GBC (Michel et al. 2010) that are expected to "transform the interaction between humanities scholars and corpus linguists" (McEnery 2015: 1–2), has not changed this situation. This is due to the fact, as mentioned in Section 1.3.5, that it is a principled problem that cannot be solved by increasing the number of texts included in a corpus, no matter whether it is synchronic or diachronic. From a statistical point of view, this means that the records cannot be used to draw inferences about changes or the evolution of the language as a whole.

However, a similar argument could be presented for the field of archeology, because it is only possible to "work with what we are given by the chance of preservation and discovery" (Jackes 2011: 107). Like archeological artefacts, I believe that diachronic corpora can still help us "improve our methods and our understanding of the past" (Jackes 2011: 138).

This idea can be illustrated by an example: if a researcher wants to analyze linguistic change, she or he could use a national newspaper or magazine such as the German "Die Zeit" or the American news magazine "TIME Magazine" (Davies 2007). This is a convenient strategy, because access to the data is fairly easy. Both journals have a long temporal coverage and due to their size allow "for accurate analysis of linguistic change across the decades." (CoRD 2015). In principle, I have no objections against using such a convenience sample, however: "statistical inference with convenience samples is a risky business" (Berk and Freedman 2003: 17), because "[a]n investigator who assumes

that a convenience sample is like a random sample seeks to obtain the benefits without the costs—just on the basis of assumptions." (Berk and Freedman 2003: 15). Size by no means guarantees representativeness and the idea that a corpus that consists of journal-ese texts is a random sample of the language as a whole is more than doubtful (cf. Section 1.3). On the other side, this in no way implies that the data collection cannot be used to find out interesting facts about linguistic change, as long as the produced results are categorized accordingly: preliminary evidence that can *improve our understanding* of linguistic change and can be used to stimulate further research.

Against this backdrop, a data-driven method, which automatically extracts word types from diachronic corpora that have undergone the most pronounced change in frequency in a given period of time, will be presented in the following. It will be shown that this method, first developed by Kilgarriff (2001; 2012) to compare synchronic corpora, combined with statistical methods from time series analysis is capable of finding meaningful patterns and relationships in diachronic corpora that can help to *improve our understanding* of linguistic change.

## **2.1 Data**

Corpus data from three different sources are used in this chapter.

(A) The first resource is a set of three unlemmatised frequency lists compiled by Kilgarriff (1997) on the basis of the BNC. One represents the written part, which comprises 89.7 million tokens. The other two frequency lists represent the spoken part of the BNC. The first of these covers 'context-governed' spoken material (e.g. lectures, meetings, news commentaries) and consists of 6.2 million tokens. The second list covers

'demographic' spoken material (i.e. recorded daily conversations) and consists of 4.2M tokens. The lists are freely available online (Kilgarriff 2014).

(B) Secondly, I use unigram data from the COHA, which contains 400 million tokens from the period between 1810 and 2009. It contains all unique word strings that occur at least three times in total. The data are freely available online (Davies 2010a).

(C) Thirdly, I use unigram data from the GBC, made available by Michel et al. (2010). For this study, the datasets of Version 2 (July 2012) of the following languages were used (including two varieties of English): American English, British English, French, German, Italian and Spanish. All unigram corpora share the same basic structure, in which the first column is the string variable for the word, the second variable contains the word-class (POS) information as described in Lin et al. (2012) and the third column contains the match count for one particular year (e.g. match1899). The data contains all unique word strings that occur at least 40 times in total<sup>11</sup>. The data are freely available online (Culturomics 2014).

Corpus size varies considerably for the synchronic data (A) and strongly increases as a function of time for the diachronic data (B and C). To avoid a potential systematic bias (Tweedie and Baayen 1998), an efficient and computationally cheap solution is to draw random samples of 1,000,000 tokens from the data by performing a binomial split for each corpus (as suggested by Piantadosi 2014). For each word type  $w$ , this procedure returns binomial  $(n_{wc}, p_c)$  random variates, where  $n_{wc}$  is the raw token frequency of the word type  $w$  in the corpus  $c$  and  $p$  is the probability of success, which is given as:

---

<sup>11</sup> Some problems that result from this truncation strategy are discussed in Koplenig and Müller-Spitzer (under review).

$(1,000,000 + 10,000)/N_c$ , where  $N_c$  is the corpus size of corpus  $c$ .<sup>12</sup> The resulting corpora of 1,000,000 tokens are what Tweedie & Baayen (1998) would call fully randomised samples of all texts in a given corpus or in a given year.<sup>13</sup>

## 2.2 Measuring similarity in synchrony

In this section, the mathematical background of the method to measure (dis-)similarity in synchrony, which will be adapted and extended to measure similarity in diachrony in the following sections, is discussed and put into perspective in regard to other measures of statistical divergence.

Kilgarriff (2001) discusses and evaluates several methods for the measurement of synchronic corpus similarity. All measures only need frequency vectors as input, using the token frequency of each orthographic word form. To compare two different corpora  $c_1$  and  $c_2$ , the individual token frequencies are first aggregated to generate a new corpus  $u$ . For his analysis, Kilgarriff (2001: 253) then uses the 500 most frequent words of  $u$ , but notes that this is done for convenience only, there is no statistical reason for this choice. For the analyses in this thesis, all words in  $u$  are used to calculate the similarity measure. In the experiment presented below, I will demonstrate that the correlation between a similarity measure based on the top 500 words and a measure based on all words is generally very high.

---

<sup>12</sup> Since this process is random *per definition*, instead of using 1,000,000 as the nominator for  $N_c$ , 1,010,000 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 were "thrown" in an urn from which 1,000,000 million tokens were then drawn randomly.

<sup>13</sup> 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). 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.

Kilgarriff (2001) works out a  $\chi^2$ -based measure. In his analysis, this measure works best for the comparison of different corpora. For each word  $w_i$  and each corpus  $c$ , an expected frequency  $e_i$  based on the union of the two corpora  $u$  is calculated using the following formula:

$$e_{i,c} = p_{i,u} \cdot N_c \quad (1)$$

where  $p_{i,u}$  is the relative frequency of word  $i$  in the union of the two corpora  $u$  and  $N_c$  is the corpus size of corpus  $c$ . Then the partial contribution  $\chi_{i,c}^2$  to the 'grand total' similarity  $\chi^2$  of word  $w_{i,c}$  with an observed frequency  $o_{i,c}$  in corpus  $c$  is defined as:

$$\chi_{i,c}^2 = \frac{(o_{i,c} - e_{i,c})^2}{e_{i,c}} \quad (2)$$

In this paper, the partial value will be used to identify the words that are most important for measuring synchronic and especially diachronic corpus similarity.

The similarity  $\chi^2$  between two corpora is then just the sum of all partial values where  $v$  is the vocabulary size of  $u$ .

$$\chi^2 = \sum_{i=1}^v \chi_{i,c}^2 \quad (3)$$

The rationale of this procedure is very intuitive: if  $c_1$  and  $c_2$  are very similar, then the distribution of token frequencies should also—apart from random fluctuations—be very similar. Therefore, the individual deviations from the expected frequencies as calculated in (1) will tend to be very small. If, for example, two identical frequency lists are being compared,  $\chi^2$  will be zero. If however  $c_1$  and  $c_2$  are very dissimilar, the expected frequencies for some words will be quite different from the observed frequencies and the squared sum of those differences will be very high. The two plots of (A) in Figure 4 visualize this idea: To compare two different corpora, the frequency differences be-

tween the word types are used as a proxy for general similarity. As can be seen on the left side of plot A, the distributions of token frequencies are – apart from random fluctuations – very similar, the lines are almost parallel. The right side of plot A shows that if  $c_1$  and  $c_2$  are very dissimilar, the expected frequencies of some words are quite different from their observed frequencies, and the squared sums of those differences are very high. For diachronic data (plot B), one corpus is analyzed at different moments in time. Thus, the method helps to identify the word types that underwent the most pronounced changes in frequency (left side). In addition to that, the method can be used to detect correlated changes (right side). This idea is developed in more detail in Section 2.4.

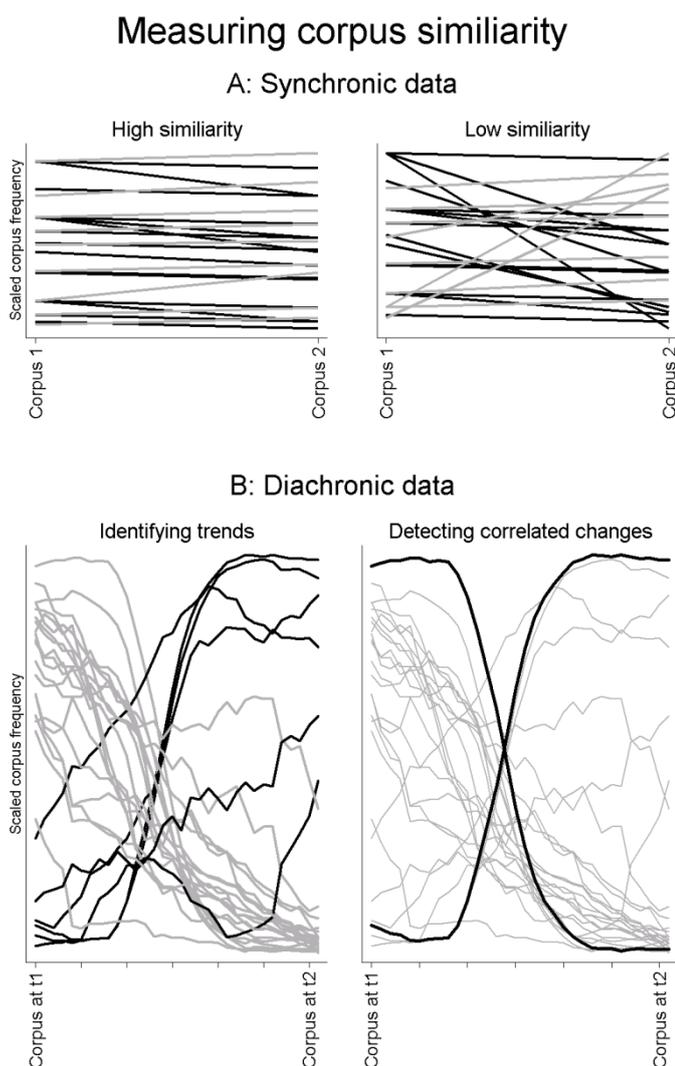
Illustrative examples of the method can also be found in Kilgarriff and Salkie (1996). Kilgarriff (2001: 255) notes that it is a desirable consequence of the approach that differences for higher-frequency words are more important in determining their individual contribution to the (dis-)similarity. If, for instance, the determiner 'the' has a very different frequency in two corpus samples, then this signals a stronger dissimilarity compared to a frequency difference for a relatively infrequent element.

Given the fact that the measure is not scale-independent (Kilgarriff 2001: 258), Cramér's  $V$  (Cramér 1946: 282) is used instead of Kilgarriff's original approach. For the comparison of two corpora,  $V$  is given as:

$$V = \sqrt{\frac{\chi^2}{N_u}} \quad (4)$$

where  $N_u$  is the corpus size of the union of the two corpora  $u$ . Cramér's  $V$  is a classical measure of the strength of association, ranging from 0 for no association to 1 for a very

strong association. For the comparison of two corpora, small values signal a very high similarity, while higher values are indicative of dissimilar corpora.



**Figure 4: Abstract visualization of the method adapted in this section.**

Compared to other measures used in his analysis, Kilgarriff (2001: 258) notes that the  $\chi^2$ -based measure is 'not rooted in a mathematical formalism' and identifies this as an area for future research. However, as a study by Endres and Schindelin (2003) suggests, the measure is, if all words are used to calculate the similarity, approximately equivalent

to the Jensen-Shannon divergence (JSD), which can be defined as (Klingenstein et al. 2014):

$$JSD(\vec{c}_1 || \vec{c}_2) = 0.5 \left[ KL \left( \vec{c}_1 \middle| \frac{\vec{c}_1 + \vec{c}_2}{2} \right) + KL \left( \vec{c}_2 \middle| \frac{\vec{c}_1 + \vec{c}_2}{2} \right) \right] \quad (5)$$

where  $\vec{c}_1$  and  $\vec{c}_2$  are the two (relative) corpus frequency vectors and  $KL$  is the Kullback-Leibler divergence, which is given as:

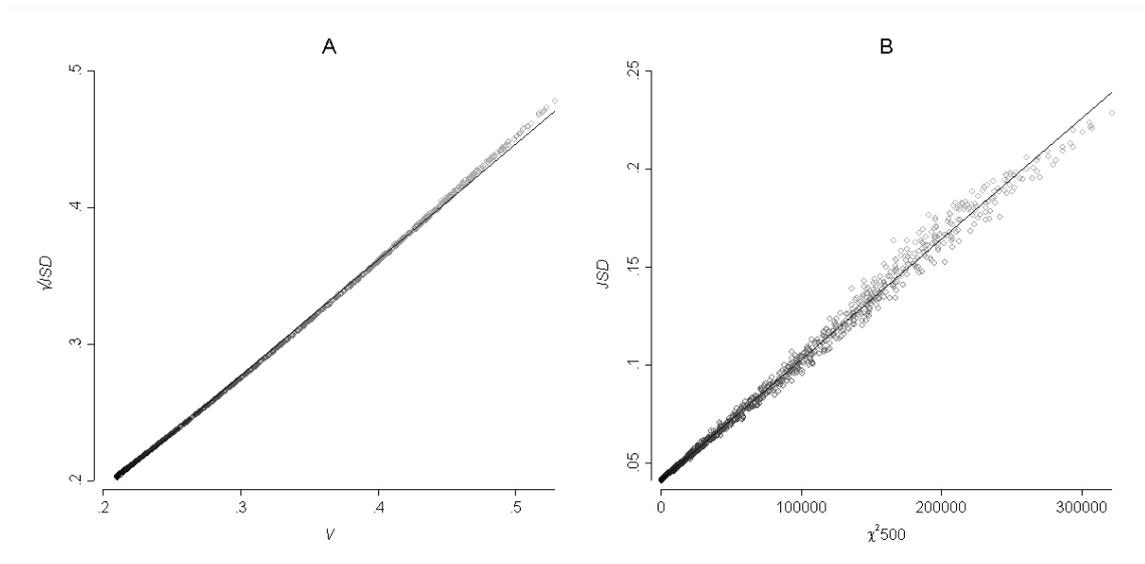
$$KL(P || Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (6)$$

where  $P(x)$  and  $Q(x)$  are the probability distributions of the two vectors. When  $P(x) = 0$  but  $Q(x) \neq 0$ , the KL-divergence is undefined, since the logarithm of zero is also undefined. For the analysis presented below (cf. Figure 5), it is stipulated that those cases are interpreted as zero in the calculation of the sum in (6), because  $\lim_{x \rightarrow 0} x \log(x) = 0$ . Bochkarev et al. (2014) demonstrate that more principled but also more complicated approaches in this context do not lead to qualitatively different results. In practice, the Jensen-Shannon divergence measures the divergence of each distribution from the mean of the two distributions (Jurafsky and Martin 2009: 700). It has already been fruitfully employed in the context of measuring stylistic influences in the evolution of literature (Hughes et al. 2012), cultural and institutional changes (Klingenstein et al. 2014), or the dynamics of lexical evolution (Bochkarev et al. 2014). Pechenick et al. (2015) use the JSD measure to quantify the changing compositions of the English GBC set with a particular attention to word types with a high partial contribution to the observed divergence (cf. Section 2.2 and Section 2.4).

To show that both  $V$  and the square root of the  $JSD$  are strongly correlated, I sampled two million tokens from the written part of the BNC (cf. Section 2.1). The resulting

frequency list is used as the union corpus  $u$ . For each word type  $w$ , a binomial split was performed, where  $n_{w,c}$  is the raw token frequency of the word type  $w$  in  $u$  and  $p$  is the probability of success. The resulting number of tokens  $f$  then represents the token frequency of  $w$  in  $c_1$ , while the difference between the total token frequency in  $u$  minus  $f$  represents the token frequency of  $w$  in  $c_2$ . The success probability  $p$  was randomly varied over the interval  $0.5 \pm r$ . To generate pairs of corpora with varying degrees of similarity,  $r$  was gradually incremented from 0.004 to 0.4. Corpus pairs with a small  $r$  should have a greater similarity than corpus pairs with a bigger  $r$ , because – on average – half of the tokens of word type  $w$  in  $u$  are classified as belonging to  $c_1$  and half of the tokens as belonging to  $c_2$ . With an increasing  $r$  however, the probability of success for each word type varies more strongly around 0.5. So, for some word types, more tokens are placed into  $c_1$ , while for other word types, most tokens are put into  $c_2$ . On average, the corpus size for each corpus remained approximately equal ( $\approx 1$  million tokens).

Using this technique, 1,000 pairs of corpora were generated and the *JSD* and the  $\chi^2$ -value for the first 500 words as suggested by Kilgarriff (2001: 253) were calculated for all pairs,  $V$ . Figure 5 demonstrates that all three measures seem to work as intended: corpus pairs with smaller values of  $r$  (indicated by darker shades of gray) are classified as being more similar (indicated by a small *JSD*, a small  $V$  and a small  $\chi^2$ -value). Corpus pairs with higher values of  $r$  (indicated by lighter shades of gray), on the other hand, are classified as being less similar.



**Figure 5: Measuring the distance between corpora with varying degrees of similarity (darker shades of gray indicate more similar corpora). Solid black lines: linear fit between the observed values (hollow circles).**

In addition to that, Plot A of Figure 5 shows that the square root of  $JSD$  and  $V$  are strongly correlated ( $\rho = 0.9996$ ) and approximately equivalent. Plot B of Figure 5 demonstrates that the  $\chi^2$ -value for the first 500 words and the  $JSD$  are also strongly related ( $\rho = 0.9972$ ). For all subsequent analyses, all words are used to calculate the similarity between two corpora as mentioned above.

### **2.3 Case study: Visualizing the difference between the two spoken parts of the BNC**

Although the main focus of this thesis is on the analysis of diachronic data, a synchronic application can be useful to illustrate the potential of the method. For this purpose, I first measured the similarity between one million token samples of the written and the two spoken parts of the BNC.

The greatest similarity is calculated for the two spoken parts ( $V = 0.384$ ). It is also in line with *a priori* expectations that the demographically sampled spoken part (*demog*) is less similar to the written sample ( $V = 0.598$ ), compared to the context-governed part (*cg*) of the BNC ( $V = 0.448$ ), because the *demog* part primarily consists of informal English conversations, while the *cg* part incorporates more informative spoken material (Burnard 2007).

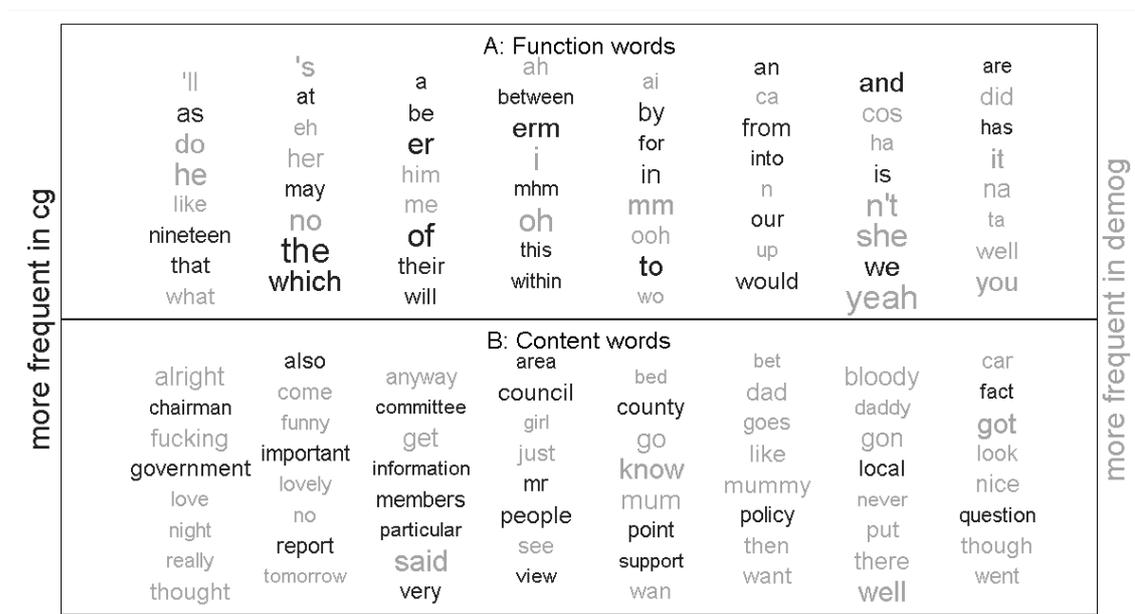
In this context, one might be interested in the main differences between these two parts of the BNC. Using formula (2) to measure the partial contribution for each word type, Figure 6 plots the elements with the most pronounced differences, showing 64 function words and 64 content words.<sup>14</sup> The size of the words is proportional to the (log of the) partial contribution of the respective word type. Lighter shades of gray indicate that the word type is more frequent in the *demog* sample, while word types in black are more frequent in the *cg* sample.

For example the determiner 'the' is the word type with the highest partial contribution ( $\chi^2 = 3276.75$ ). It has a *cg*-token frequency of 47,592 and a much lower *demog* frequency of 27,421. In reference to a classical study on genre differences (Biber and Finegan 1989), the visualization fits nicely with the fact that the *demog* part is dominated by daily conversations which can be characterized by linguistic features that are typical for interactive or involved text production, e.g. present tense verbs, contractions ('ll', 'm', 's', 'n't'), 1<sup>st</sup>- & 2<sup>nd</sup>-person pronouns (except 'we') or the pronoun 'it'. On the other side of

---

<sup>14</sup> Content words were defined as belonging to one of the following CLAWS POS category (cf. <http://www.kilgariff.co.uk/BNClists/poscodes.html>; last accessed 21.Okttober 2014): aj0, aj0-av0, aj0-nn1, aj0-vvd, aj0-vvg, ajc, ajs, av0, nn0, nn1, nn1-np0, nn1-vvb, nn1-vvg, nn2, nn2-vvz, np0, vvb, vvd, vvd-vvn, vvg, vvi, vvn, vvz. The rest of the word types were flagged as function words. The analysis was conducted separately for each category.

this continuum, the *cg* part of the BNC contains more prepositions (e.g. 'in', 'to', 'into', 'at', 'by', 'from') and more nouns (as indicated by the higher frequency of the determiners 'the' and 'a'), which are linguistic features that signal informational text production (Biber and Finegan 1989: 490–492).



**Figure 6: Visualization of the differences between the two spoken-parts of the BNC. The size of the words is proportional to the (log of the) partial contribution of the respective word type. Lighter shades of gray indicate that the word type is more frequent in the demog sample, while black word types are more frequent in the *cg* sample.**

Accordingly, an additional analysis reveals that the *cg* sample contains 15,970 different nouns, while the *demog* sample only contains 13,052 nouns. The *written* sample even contains 36,072 nouns (regular expression to count the number of nouns: `nn*`). Further analyses focusing on different word classes are possible using this approach (cf. Section 2.6).

## 2.4 Measuring corpus similarity in diachrony

The illustration presented in the last section showed how the Kilgarriff approach (2001) can be used to fruitfully measure and visualize differences between synchronic data. Since a diachronic corpus is basically just "a collection of texts that vary along the parameter of time" (Gries and Hilpert 2008: 386, cf. Section 2.), the focus will now move on to the analysis of diachronic data, using the same methodology (cf. Mota 2010 for a similar approach in this direction). Instead of comparing two different corpora, synchronic snapshots (i.e. word type vectors of token frequencies) of one diachronic corpus are compared across different successive moments in time. In this case, the similarity  $V$  between time point  $t_1$  and  $t_2$  is the square root of the sum of all partial values divided by corpus size of the union of the two synchronic snapshots. Using this methodology, small values of  $V$  indicate that in the investigated period of time, no pronounced lexical changes have taken place. If however, the frequencies of many word types have changed, the expected frequencies of those words will strongly diverge from the observed frequencies, and as a consequence, the value of  $V$  will be bigger.

Compared to the synchronic case (cf. plot A of Figure 4), using formula (2) to measure the partial contribution for each word type has a natural interpretation for diachronic data (cf. the left side of plot B of Figure 4): large partial values indicate that the respective word types underwent pronounced changes in frequency. In terms of time series analysis, this means that word types with a large partial value have a strong upward or downward trend. This property allows us to adopt a new analytical perspective, because the method can be used to detect correlated changes (cf. the right side of plot B of Figure 4), that is, it can find word types whose frequency changes are positively or nega-

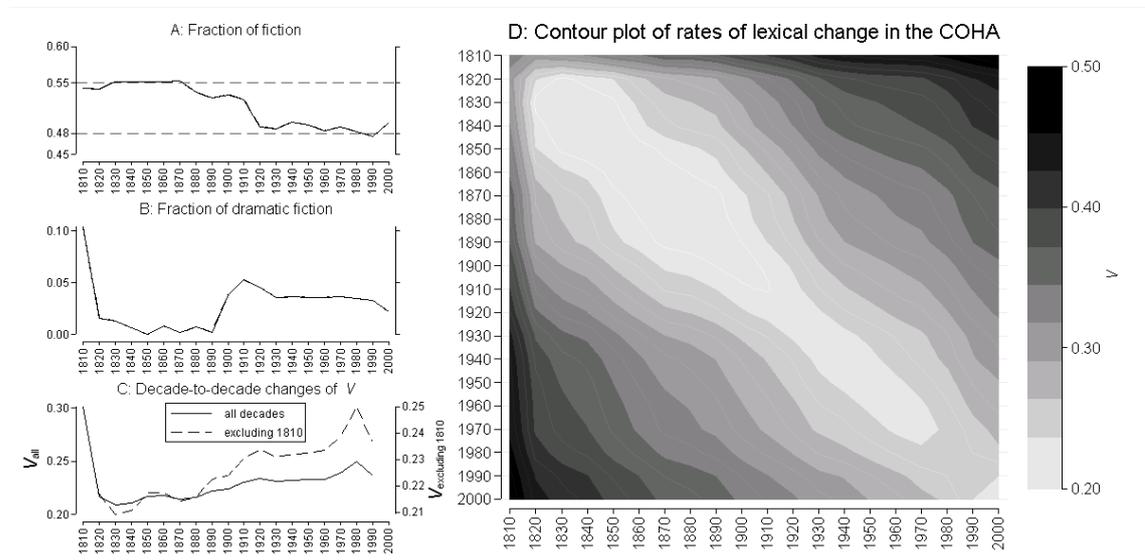
tively correlated (e.g. word types that have potentially replaced each other). This is discussed in Chapter 4 and Chapter 5.

In the next section, a case study shows that the method is useful for the identification of lexical changes. In addition, it is demonstrated in Koplenig (under review) that the method (i) is able to identify short term diachronic shifts that can be linked to historical events, (ii) helps to improve diachronic POS tagging, and (iii) complements other NLP approaches. This indicates that the approach can be fruitfully applied to the analysis of diachronic processes in linguistic change.

## 2.5 Case study: Lexical changes in the COHA

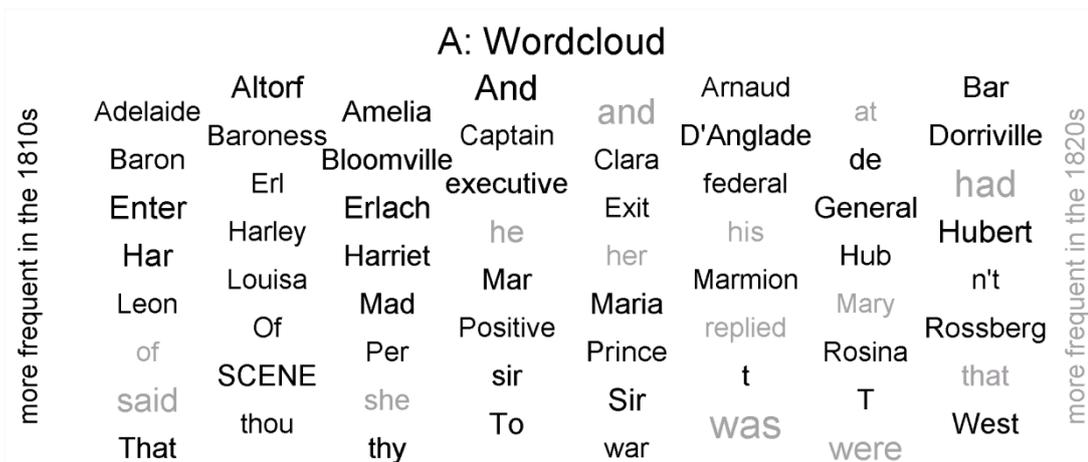
Compared to the GBC, the COHA is balanced in regard to both genre and sub-genre across decades. It is hoped that this "allows researchers to examine changes and be reasonably certain that the data reflects actual changes in the 'real world', rather than just being artifacts of a changing genre balance" (Davies 2010a). Using the information about the composition of the COHA (Davies 2010b), plot A of Figure 7 shows that the ratio of fiction is closely around 50% for each decade (48 – 55%). Plot B of Figure 7 demonstrates that the ratio of dramatic fiction varies between zero and roughly ten percent (in 1810). To measure similarity between decades,  $V$  is calculated as described above. The solid line in plot C of Figure 7 plots the value of  $V$  between the decade  $d$  and  $d+10$  for all decades (left y-axis), while the dashed line plots  $V$  excluding the 1810s (right y-axis), since the large difference between the 1810s and the 1820s would mask subsequent processes. Comparing plot B and plot C demonstrates that  $V$  seems to capture the higher fraction of dramatic fiction in the 1810s (solid line). In addition, the

dashed line also shows that  $V$  is sensitive to the fact that the ratio of dramatic fiction rises around the 1900s and remains at a similar level afterwards. The contour plot D of Figure 7 shows  $V$  between all pairs of decades.

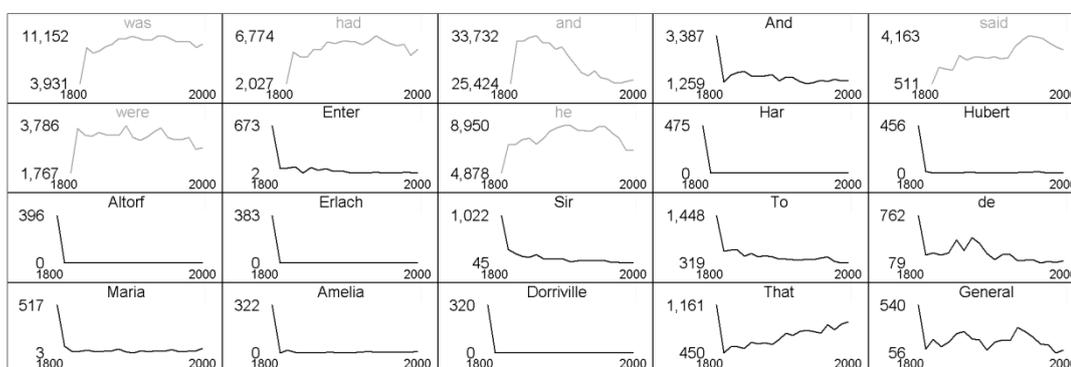


**Figure 7: Visualizing changing genre compositions of the COHA using the  $V$ -method.**

One might object that it is not clear whether the observed changes can really be attributed to the changing ratio of dramatic fiction (cf. Plot C of Figure 7). To refute this objection, Figure 8 plots the 64 most different word types between the 1810s and the 1820s. Additionally, time series for the 20 word types with the highest partial contribution are visualized as small multiples (Tufte 2001).



**B: Time series for top 20 most different**



**Figure 8: Visualization of the differences between the 1810s and the 1820 in the COHA. The size of the words is proportional to the (log of the) partial contribution of the respective word type. Lighter shades of gray indicate that the word type is more frequent in the 1820s, while black colored word types are more frequent in the 1810s. The labels on the y-axes of the small multiples show maximum and minimum values in the 1 million token samples (cf. Section 2.1).**

Figure 5 clearly shows that the higher ratio of dramatic fiction for the 1810s explains the observed dissimilarity between this decade and the other decades. Words like 'Enter', 'Exit', or 'SCENE' and proper nouns in general are typical for dramas. Using the information about the composition of the COHA further reveals that 'Marmion' is a play by J. N. Barker; 'Hub' and 'Mar' are abbreviations of names (Hubert & Marcello) in a play by I. Harby called the Gordian Knot; Harriet 'Har' Bloomville is one of the main

characters in 'The Poor Lodger' by W. C. White and if one reads 'The Tooth-Ache' by B. John, one finds out the 'de' is used to dramatize a strong French accent as in 'dis is de same vay, avec tout le monde [...] you see de pain go away for little [...]' (p. 26).

To sum up this section, the analysis indicates that the method outlined above can help to automatically identify and understand changing genre compositions in diachronic corpora.

## 2.6 Further applications

To discover "previously undetected phenomena available for further analysis" (Hilpert and Gries 2009: 398), the reader is invited to visit <http://www.owid.de/plus/lc2015/>, where it is possible to choose one of seven languages (three varieties of English [American English, British English, English Fiction], French, German, Italian and Spanish), two periods of time from 1800 to 2000 and various part of speech categories. Based on the method presented in this paper, the tool then automatically visualizes the words whose frequency changed most in the GBC for the selected time periods (similar to Figure 8).<sup>15</sup>

One example that might be interesting for cross cultural and linguistic discourse analyses can be found in the Appendix (Section 2.9). It demonstrates that the method makes it easy to replicate and extend analyses like the one presented by Pechenick et al. (2015), who argue that the English GBC have seen a rise of scientific and technical literature throughout the second half of the 20<sup>th</sup> century. Indeed, Figure 9 - Figure 15 clearly demonstrate that this seems to be not only true for the two English corpora [American and British English] but also for the French, the German, the Italian and the

---

<sup>15</sup> I would like to thank Frank Michaelis for programming the graphical user interface.

Spanish corpora, as can be seen by an increase of the respective vocabulary (e. g. 'information', 'model', 'data', 'research', 'modèle', 'recherche', 'approche', 'Daten', 'Modell', 'approccio', 'modello', 'ricerca', 'modelo', 'tecnología'). The visualization technique used in this paper also makes it possible to further refine the analysis of Pechenick et al. (2015) who argue that there was a rise of the two *science-related* terms 'percent' and 'Figure'. Figure 9 reveals that 'percent' was written 'per cent' in earlier periods (cf. Koplenig 2015a) and 'Figure' was abbreviated 'Fig.' or 'F.' in the American English GBC.

It is also worth pointing out that there is a rise of gender related vocabulary ('women', 'gender', 'femmes', 'Frauen', 'donne', 'mujeres', cf. Figure 7 - Figure 15) in all six GBC corpora. This could stimulate further research in historical discourse analysis.

## 2.7 Summary of the studies

As written at the end of Section 1, I argue that diachronic corpus data as non-representative (in a statistical sense) language samples can nevertheless be used to improve our understanding of linguistic change and stimulate further research. At the core of this approach is the key idea that it is essential to work with the corpus data: in order to understand potential confounding variables, or put differently, in order to disentangle what Szmrecsanyi (2015) calls environmental and actual linguistic change, it is important (i) to use all available sources of information in the data and (ii) inspect the corpus data as much as possible. While I focus on (i) in the first paper (cf. Section 3), both the second paper (cf. Section 4) and the third paper (cf. Section 5) demonstrate the importance of (ii).

When the GBC were made available for public use (Culturomics 2014), many researchers hoped that this vast amount of data would enable them to study linguistic and cultur-

al change with unprecedented accuracy. However, to avoid breaking any copyright laws, the datasets are not accompanied by any metadata regarding the texts the corpora consist of. This absence of sufficient metadata is exactly what can and what has been criticized about studies based on the GBC data by various linguists, published mostly in blogs (Lieberman 2012; Jockers 2010; Jockers 2013; Underwood 2012). Since most of the arguments that were put forward were rather vaguely and focused mainly on pointing out that the lack of metadata generally is problematic, one could again refer to Berk and Freedman (2003: 4) and argue that *handwaving* is also *inadequate* in this context. To counter this claim, one needs evidence why and when the lack of metadata actually matters. To this end, the developed method (cf. Section 2.4) is used in the first paper (cf. Section 3) to show why the lack of metadata matters: I chose the example of measuring censorship in Nazi Germany, which received widespread attention and was published in a paper that accompanied the release of the GBC (Michel et al. 2010). I show that without proper metadata, it is unclear whether the results actually reflect any kind of censorship at all. On the contrary, the presented results support the argument that the German GBC was strongly biased towards volumes published in Switzerland during WWII. Collectively, the findings imply that observed changes in this period of time can only be linked directly to World War II to a certain extent. On a general level, the results of this study demonstrate that the importance of metadata cannot be underestimated: the availability of metadata is not just a nice add-on, but a powerful source of information for the digital humanities. For all kinds of research in this context – to rephrase a quote by Biber (1998, p. 249) – it is important to realize that size cannot make up for a lack of metadata.

The second paper (cf. Section 4) is of more methodological nature. Its main point is to demonstrate why a quantitative 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) 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: 14). This re-analysis shows that apparent relationships like this are the result of misspecified models whose validity has to be questioned. Given the fact that in the recent past, several studies were published— some in journals with a good reputation – that seem to suffer from exactly this problem (Bentley et al. 2014; Hills and Adelman 2015; Hills et al. 2015; Frimer et al. 2015; Twenge et al. 2012; Zeng and Greenfield 2015), I discussed the cause of the misspecification and its profound consequences in another recent publication, together with Carolin Müller- Spitzer (Koplenig and Müller-Spitzer, under review).

The third paper (cf. Section 5) starts with the review of a very famous observation that word frequency distributions tend to be distributed in a special mathematical way (Zipf 1935; Zipf 2012) that have some interesting consequences and that can be used to understand different linguistic phenomena (Bentz, Kiela, et al. 2014; Bentz, Verkerk, et al. 2014; Yang 2013; Baixeries et al. 2013). In the paper, the GBC are used for a cross-linguistic comparison. For this purpose, the statistical properties of this well-known fact about languages are used to measure a second – equally well-known – fact about languages, namely that languages are constantly changing on all fundamental levels (Labov 1994). Referring to Berk and Freedman (2003) again, the idea of this paper was

that while taking for granted that the GBC are not representative in any sense, they still constitute a valuable sample of the written language record: they can be used in order to test if relationships based on text collections for one language, also hold for other languages.<sup>16</sup> It is demonstrated that diachronic changes of the parameters of the distribution can be used to quantify and visualize important aspects of linguistic change. On the other hand, the analysis also revealed that there are important cross-linguistic differences. On this basis, one can argue that the statistical properties of word frequency distributions 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 stylistometric measures to fully understand the factors that actually drive those changes.

## 2.8 References

- Angrist, Joshua D. & Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton NJ: Princeton University Press.
- Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27.
- Arppe, Antti & Juhani Järvi. 2007a. Take empiricism seriously! - In support of methodological diversity in linguistics [Commentary of Geoffrey Sampson 2007. Grammar without Grammaticality.]. *Corpus Linguistics and Linguistic Theory* 3(1). 99–109.

---

<sup>16</sup> If an astute reader would ask me why some of the results presented in this paper include coefficients of significance, I would give two different answers and leave it to her or him to decide which alternative to pick:

- i) Compared to cross-sectional data, the concept of a population from which the time series is drawn as a sample is different because the population mean does not necessarily exist at all (Chatfield 2004:chap. 4). Therefore the analysis only focusses on relating different time-series that are realizations from the population of all potential time series with each other.
- ii) If I would write this paper again today, I would not use *p*-values again.

- Arppe, Antti & Juhani Järvi. 2007b. Every method counts - Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.
- Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5. 436–461.
- Baixeries, Jaume, Brita Elvevåg & Ramon Ferrer-i-Cancho. 2013. The Evolution of the Exponent of Zipf's Law in Language Ontogeny. (Ed.) Satoru Hayasaka. *PLoS ONE* 8(3). e53227. doi:10.1371/journal.pone.0053227.
- Baroni, Marco & Stefan Evert. 2009. Statistical methods for corpus exploitation. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 777–802. Berlin: De Gruyter Mouton.
- Beckett, Sean. 2013. *Introduction to time series using Stata*. 1st ed. College Station, Tex: Stata Press.
- Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod & Vasileios Lampos. 2014. Books Average Previous Decade of Economic Misery. (Ed.) Matjaž Perc. *PLoS ONE* 9(1). e83147. doi:10.1371/journal.pone.0083147.
- Bentz, Christian, Douwe Kiela, Felix Hill & Paula Buttery. 2014. 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. 2014. 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](http://www.christianbentz.de/Papers/Bentz%20et%20al.%20(submitted)%20Adaptive%20Languages.pdf) (8 September, 2014).
- Berk, Richard A. & David A. Freedman. 2003. Statistical assumptions as empirical commitments. In Sheldon L. Messinger, Thomas G. Blomberg & Stanley Cohen (eds.), *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. New York: Aldine de Gruyter.  
<http://www.stat.berkeley.edu/~census/berk2.pdf> (15 June, 2015).
- Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4). 243–257. doi:10.1093/lc/8.4.243 (30 March, 2015).
- Biber, Douglas. 1998. *Corpus linguistics: investigating language structure and use*. (Cambridge Approaches to Linguistics). Cambridge ; 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 (1 July, 2014).
- Bochkarev, Vladimir, Valery Solovyev & Sören Wichmann. 2014. Universals versus historical contingencies in lexical evolution.  
<http://wwwstaff.eva.mpg.de/%7Ew Wichmann/LexEvolUploaded.pdf> (12 June, 2014).

- Brezina, Vaclav & Miriam Meyerhoff. 2014. Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1). 1–28. doi:10.1075/ijcl.19.1.01bre.
- Burnard, Lou (ed.). 2007. [bnc] British National Corpus. <http://www.natcorp.ox.ac.uk/docs/URG/> (21 October, 2014).
- Campbell, Donald T. & Julia C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Skokie, Ill: Rand McNally.
- Caruana-Galizia, Paul. 2015. Politics and the German language: Testing Orwell’s hypothesis using the Google N-Gram corpus. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqv011. <http://dsh.oxfordjournals.org/cgi/doi/10.1093/llc/fqv011> (15 April, 2015).
- Chatfield, Christopher. 2004. *The analysis of time series: an introduction*. 6th ed. (Texts in Statistical Science). Boca Raton, FL: Chapman & Hall/CRC.
- Chomsky, Noam. 1986. *Knowledge of language: its nature, origin, and use*. (Convergence). New York: Praeger.
- Chomsky, Noam. 1988. *Language and problems of knowledge the Managua lectures*. (Current Studies in Linguistics Series 16). Cambridge, Mass.: MIT Press.
- CoRD. 2015. CoRD | The TIME Magazine corpus (TIME). <http://www.helsinki.fi/varieng/CoRD/corpora/TIME/> (18 June, 2015).
- Cramé, Harald. 1946. *Mathematical methods of statistics*. Princeton: Princeton University Press.
- Culturomics. 2014. [www.culturomics.org](http://www.culturomics.org). [www.culturomics.org](http://www.culturomics.org). <http://www.culturomics.org/> (8 September, 2014).
- Davies, Mark. 2007. TIME Magazine corpus: 100 million words, 1820s-2000s. <http://corpus.byu.edu/time> (16 October, 2014).
- Davies, Mark. 2010a. The Corpus of Historical American English: 400 million words, 1810-2009. <http://corpus.byu.edu/coha/> (16 October, 2014).
- Davies, Mark. 2010b. The Corpus of Historical American English: COMPOSITION OF THE CORPUS. <http://corpus.byu.edu/coha/files/cohaTexts.xls> (24 October, 2014).
- Deppermann, Arnulf & Martin Hartung. 2011. Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch“ (FOLK) am Institut für Deutsche Sprache (Mannheim). In Ekkehard Felder, Marcus Müller & Friedemann Vogel (eds.), *Korpuspragmatik*. Berlin, Boston: DE GRUYTER. <http://www.degruyter.com/view/books/9783110269574/9783110269574.415/9783110269574.415.xml> (10 June, 2015).
- Diekmann, Andreas. 2002. *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. 8th ed. Reinbek: Rowohlt Taschenbuch Verlag.
- Durrell, Martin. 2015. “Representativeness”, “Bad Data”, and legitimate expectations. What can an electronic historical corpus tell us that we didn’t actually know al-

- ready (and how)? In Jost Gippert & Ralf Gehrke (eds.), *Historical corpora: challenges and perspectives*, 13–33. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5). Tübingen: Narr.
- Ellenberg, Jordan. 2014. The Summer's Most Unread Book Is.... *The Wall Street Journal*. <http://www.wsj.com/articles/the-summers-most-unread-book-is-1404417569> (11 June, 2015).
- Endres, D.M. & J.E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* 49(7). 1858–1860. doi:10.1109/TIT.2003.813506 (17 October, 2014).
- Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–190. (19 March, 2014).
- Frimer, Jeremy A., Karl Aquino, Jochen E. Gebauer, Luke (Lei) Zhu & Harrison Oakes. 2015. A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences* 112(21). 6591–6594. doi:10.1073/pnas.1500355112.
- Gilquin, Gaëtanelle. 2008. What You Think Ain't What You Get: Highly polysemous verbs in mind and language. In Guillaume Desgulier, Jean-Baptiste Guignard & Jean Rémi Lapaire (eds.), *Du fait grammatical au fait cognitif. From Gram to Mind*, vol. 2. Pessace: Presses Universitaires de Bordeaux.
- Gilquin, Gaëtanelle & Stefan Th. Gries. 2009. Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.
- Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2). doi:10.1515/cllt.2005.1.2.277. <http://www.degruyter.com/view/j/cllt.2005.1.issue-2/cllt.2005.1.2.277/cllt.2005.1.2.277.xml> (28 May, 2015).
- Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125. doi:10.3366/cor.2015.0068.
- Gries, Stefan Th., Beate Hampe & Döris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.
- Gries, Stefan Th., Beate Hampe & Döris Schönefeld. 2010. Converging evidence II: more on the association of verbs and constructions. In John Newman & Sally Rice (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: CSLI.
- Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3(1). 59–81.
- Gries, Stefan Thomas. 2006. Introduction. In Stefan Thomas Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 1–18. (Trends in Linguistics. Studies and Monographs 172). Berlin ; New York: Mouton de Gruyter.

- Hills, Thomas, Eugenio Protio & Daniel Sgroi. 2015. Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books. IZA Discussion Paper No. 9195. Bonn, ms. <http://ftp.iza.org/dp9195.pdf> (13 August, 2015).
- Hills, Thomas T. & James S. Adelman. 2015. Recent evolution of learnability in American English from 1800 to 2000. *Cognition* 143. 87–92. doi:10.1016/j.cognition.2015.06.009.
- 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/lc/fqn012 (13 January, 2015).
- Hughes, James M., Nicholas J. Foti, David C. Krakauer & Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences* 109(20). 7682–7686. doi:10.1073/pnas.1115407109 (10 March, 2014).
- Hunston, Susan. 2010. *Corpora in applied linguistics*. 7. print. (The Cambridge Applied Linguistics Series). Cambridge: Cambridge Univ. Press.
- Jacks, Mary. 2011. Representativeness and Bias in Archaeological Skeletal Samples. In Sabrina C. Agarwal & Bonnie A. Glencross (eds.), *Social Bioarchaeology*, 107–146. Oxford, UK: Wiley-Blackwell. <http://doi.wiley.com/10.1002/9781444390537.ch5> (16 October, 2014).
- Jann, Ben. 2005. *Einführung in die Statistik*. München; Wien: Oldenbourg.
- Jockers, Matthew Lee. 2010. Unigrams, and bigrams, and trigrams, oh my. [Http://www.matthewjockers.net](http://www.matthewjockers.net). <http://www.matthewjockers.net/2010/12/22/unigrams-and-bigrams-and-trigrams-oh-my/> (10 March, 2014).
- Jockers, Matthew Lee. 2013. *Macroanalysis: digital methods and literary history*. (Topics in the Digital Humanities). Urbana: University of Illinois Press.
- Jurafsky, Daniel & James H Martin. 2009. *Speech and Language processing: an introduction to natural language processing, computational Linguistics, and speech recognition*. Upper Saddle River: Pearson Education (US).
- Kellehear, Allan. 1993. *The Unobtrusive Researcher: A guide to methods*. St. Leonards, NSW: Allen & Unwin Pty LTD.
- Kertész, András & Csilla Rákosi (eds.). 2008. *New approaches to linguistic evidence: pilot studies = Neue Ansätze zu linguistischer Evidenz: Pilotstudien*. (MetaLinguistica v. 22). Frankfurt ; New York: Peter Lang.
- 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 (19 May, 2014).
- Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). doi:10.1515/cllt.2005.1.2.263.

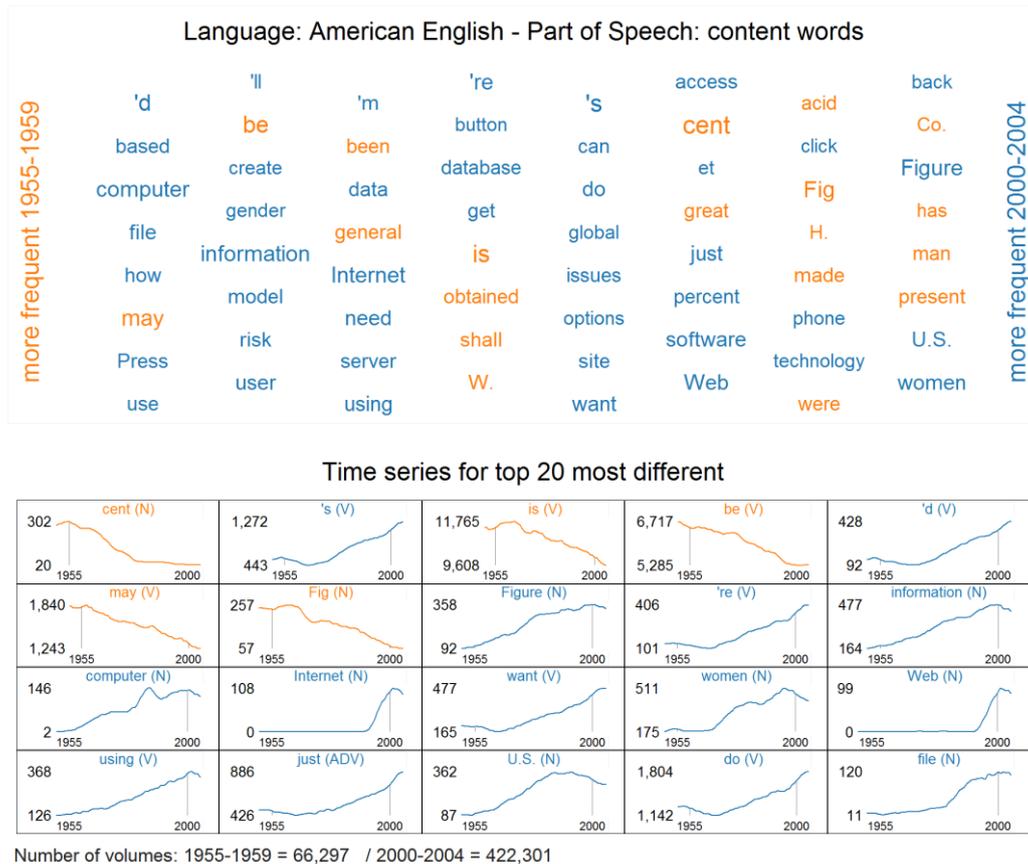
- <http://www.degruyter.com/view/j/cllt.2005.1.issue-2/cllt.2005.1.2.263/cllt.2005.1.2.263.xml> (2 May, 2014).
- Kilgarriff, Adam. 2012. Getting to Know Your Corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, Speech and Dialogue*, vol. 7499, 3–15. Berlin, Heidelberg: Springer Berlin Heidelberg.  
[http://link.springer.com/10.1007/978-3-642-32790-2\\_1](http://link.springer.com/10.1007/978-3-642-32790-2_1) (2 February, 2015).
- Kilgarriff, Adam. 2014. Read-me for Kilgarriff's BNC word frequency lists.  
<http://www.kilgarriff.co.uk/bnc-readme.html> (20 October, 2014).
- Kilgarriff, Adam & Raphael Salkie. 1996. Corpus Similarity and Homogeneity via Word Frequency. *Euralex '96 proceedings: papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden. 1. 1.*, 121–130. Göteborg: Univ., Dep. of Swedish.
- Klingenstein, S., T. Hitchcock & S. DeDeo. 2014. The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences* 111(26). 9419–9424. doi:10.1073/pnas.1405984111 (17 October, 2014).
- Köhler, Reinhard. 2005. Korpuslinguistik - zu wissenschaftstheoretischer Grundlagen und methodologischen Perspektiven. *LDV Forum* 20(2). 1–16.
- Kohnen, Thomas. 2007. From Helsinki through the centuries: The design and development of English diachronic corpora." In: Towards Multimedia in Corpus Studies. In Päivi Phata, Irma Taavitsainen, Terttu Nevalainen & Jukka Tyrkkö (eds.), *Helsinki: Research Unit for Variation, Contacts and Change in English*. (Studies in Language Variation, Contacts and Change in English 2).  
<http://www.helsinki.fi/varieng/journal/volumes/02/kohnen> (5 October, 2014).
- Koplenig, Alexander. under review. A data-driven method to identify (correlated) changes in chronological corpora. *Journal of Quantitative Linguistics*, ms.
- Koplenig, Alexander. 2015a. 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* 0(0). doi:10.1515/cllt-2014-0049. <http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2014-0049/cllt-2014-0049.xml> (19 April, 2015).
- Koplenig, Alexander. 2015b. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digital Scholarship in the Humanities*. fqv030. doi:10.1093/llc/fqv030.
- Koplenig, Alexander. 2015c. The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII. *Digital Scholarship in the Humanities*. fqv037. doi:10.1093/llc/fqv037.
- Koplenig, Alexander & Carolin Müller-Spitzer. under review. Population size predicts lexical diversity, but so does the mean sea level – one problem in the analysis of temporal data. <http://hdl.handle.net/10932/00-02AC-21D1-CCC0-0201-B>.
- Labov, William. 1994. *Principles of linguistic change*. (Language in Society 20). Oxford, UK ; Cambridge [Mass.]: Blackwell.

- Leech, Geoffrey. 1991. The state of the art in corpus linguistics. In Jan Svartvik, Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: studies in honour of Jan Svartvik*, 8–29. London ; New York: Longman.
- Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (eds.), *Corpus Linguistics and the Web*, 133–149. Rodopi.
- Liberman, Mark. 2012. Textual narcissism. *Language Log*.  
<http://languagelog.ldc.upenn.edu/nll/?p=4069> (10 March, 2014).
- Lijffijt, Jeffrey, Tertti Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamaki & Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu064.  
<http://dsh.oxfordjournals.org/cgi/doi/10.1093/llc/fqu064> (22 April, 2015).
- 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.
- Mandera, Paweł, Emmanuel Keuleers & Marc Brysbaert. 2015. How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*. 1–20. doi:10.1080/17470218.2014.988735 (22 April, 2015).
- McEnery, Tony. 2015. Editorial. *Corpora* 10(1). 1–3. doi:10.3366/cor.2015.0063 (22 April, 2015).
- McEnery, Tony & Andrew Wilson. 1996. *Corpus linguistics*. (Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University Press.
- McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: an advanced resource book*. London; New York: Routledge.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Verses, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(14). 176–182 [online pre-print: 1–12]. doi:10.1126/science.1199644.
- Mota, Cristina. 2010. Journalistic Corpus Similarity over Time. In Stefan Thomas Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus-linguistic applications: current studies, new directions*, 67–84. (Language and Computers no. 71). Amsterdam ; New York: Rodopi.
- Oakes, Michael P. 1998. *Statistics for corpus linguistics*. (Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University Press.
- Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press.
- Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. <http://arxiv.org/abs/1501.00960>.

- Perkuhn, Rainer, Holger Keibel & Marc Kupietz. 2012. *Korpuslinguistik*. Paderborn: Fink.
- 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> (2 May, 2014).
- Popper, Karl. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford Univ Press.
- Rieger, Burghard. 1979. Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In Henning Bergenholtz & Burkhard Schaefer (eds.), *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*, 52–70. (Monographien Linguistik Und Kommunikationswissenschaft 39). Königssteing/ Taunus: Scriptor. <http://www.uni-trier.de/fileadmin/fb2/LDV/Rieger/Publicationen/Aufsaeetze/79/rub79.html>.
- Schmid, Hans-Jörg. 2010. Does frequency in text instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, 101–133. Berlin, New York: de Gruyter.
- Schönefeld, Doris. 2011. Introduction. On evidence and the convergence of evidence in linguistic research. In Doris Schönefeld (ed.), *Converging evidence: methodological and theoretical issues for linguistic research*, 1–31. (Human Cognitive Processing v. 33). Amsterdam ; Philadelphia: John Benjamins Pub. Co.
- Schütze, Carson T. 1996. *The Empirical Base of Linguistics*. Chicago: The University of Chicago Press.
- Shang, Aijing, Karin Huwiler-Müntener, Linda Nartey, Peter Jüni, Stephan Dörig, Jonathan A C Sterne, Daniel Pewsner & Matthias Egger. 2005. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 366. 726–732.
- Szmrecsanyi, Benedikt. 2015. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 0(0). doi:10.1515/cllt-2015-0068. <http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2015-0068/cllt-2015-0068.xml> (28 December, 2015).
- Taylor, Jonathan & Robert J. Tibshirani. 2015. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*. 201507583. doi:10.1073/pnas.1507583112.
- Thompson, John. 2014. *Bayesian analysis with Stata*. First edition. College Station, Texas: Stata Press.
- Trochim, William. 2006. Design. *Research Methods Knowledge Base*. <http://www.socialresearchmethods.net/kb/design.php> (14 September, 2011).
- Tufte, Edward R. 2001. *The visual display of quantitative information*. 2nd ed. Cheshire, Conn: Graphics Press.

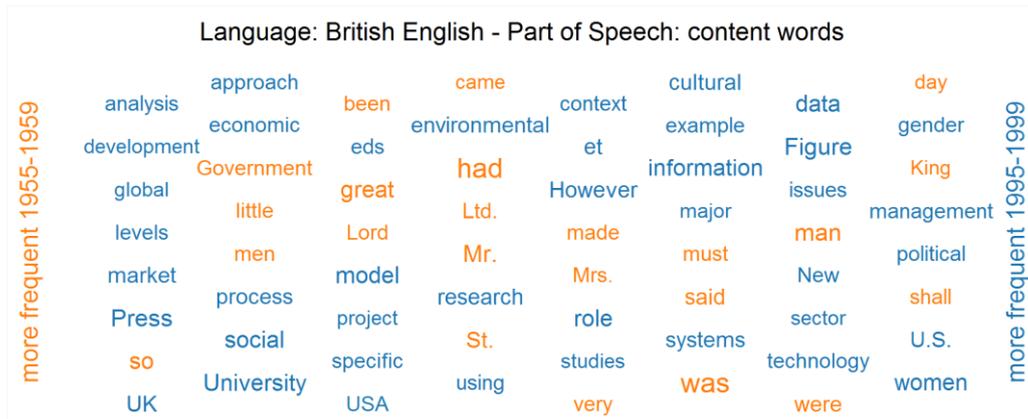
- Tweedie, Fiona J. & R. Harald Baayen. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32(5). 323–352.
- Twenge, Jean M., W. Keith Campbell & Brittany Gentile. 2012. Male and Female Pronoun Use in U.S. Books Reflects Women’s Status, 1900–2008. *Sex Roles* 67(9-10). 488–493. doi:10.1007/s11199-012-0194-7.
- Underwood, Ted. 2012. ngrams | The Stone and the Shell. *The Stone and the Shell - Historical questions raised by a quantitative approach to language*. <http://tedunderwood.com/category/ngrams/> (10 March, 2014).
- Váradí, Tamás. 2001. The Linguistic Relevance of Corpus Linguistics. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie & Shereen Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 conference, Lancaster University (UK), 29 March - 2 April 2001*, 587–593. Lancaster: Lancaster University.
- Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in Linguistic Argumentation. *Lingua* 114(11). 1481–1496.
- Wiechmann, Daniel. 2008. On the Computation of Collocation Strength. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.
- Yang, Charles. 2013. Ontogeny and phylogeny of language. *PNAS* 110(16). 6324–6327.
- Zeng, Rong & Patricia M. Greenfield. 2015. Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values: CULTURAL EVOLUTION IN CHINA. *International Journal of Psychology* 50(1). 47–55. doi:10.1002/ijop.12125.
- 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.

## 2.9 Appendix<sup>17</sup>

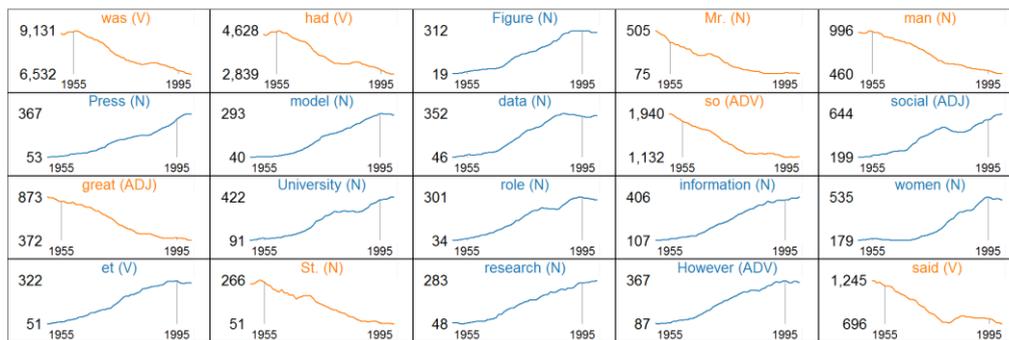


**Figure 9: Visualization of the differences between the time spans 1955-1959 and 2000-2004 in the American English GBC**

<sup>17</sup> Source of Figure 9 – Figure 15 : <http://www.owid.de/plus/lc2015/>, last accessed 02/15/2016.

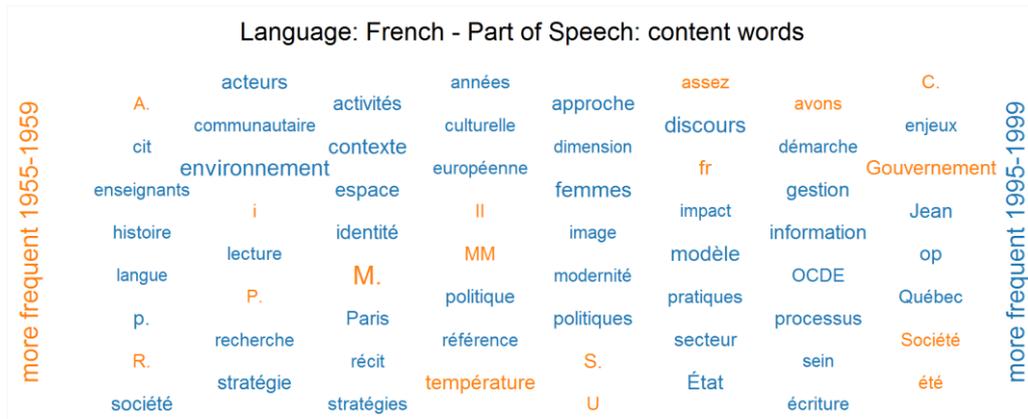


Time series for top 20 most different

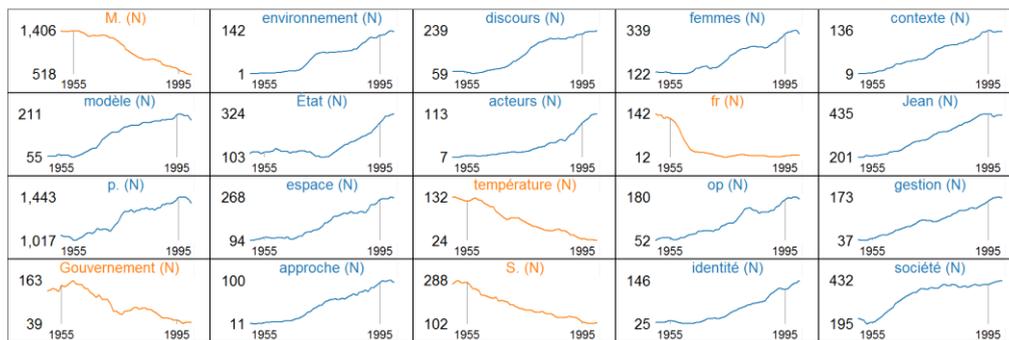


Number of volumes: 1955-1959 = 20,221 / 1995-1999 = 69,735

Figure 10: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the British English GBC



Time series for top 20 most different

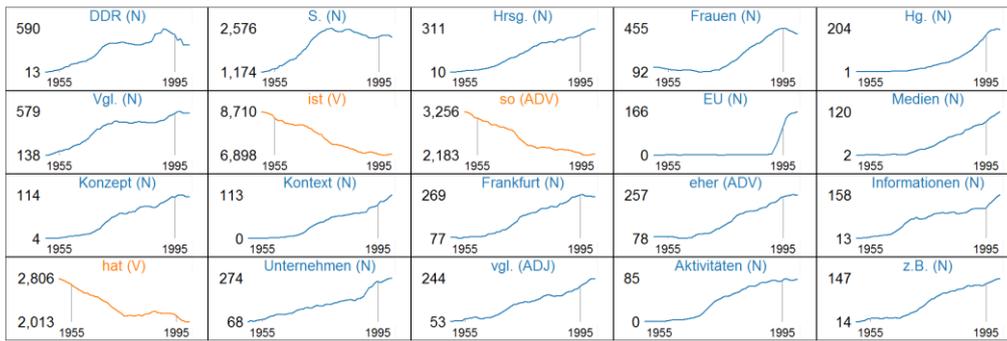


Number of volumes: 1955-1959 = 15,169 / 1995-1999 = 42,375

**Figure 11: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the French GBC**



Time series for top 20 most different

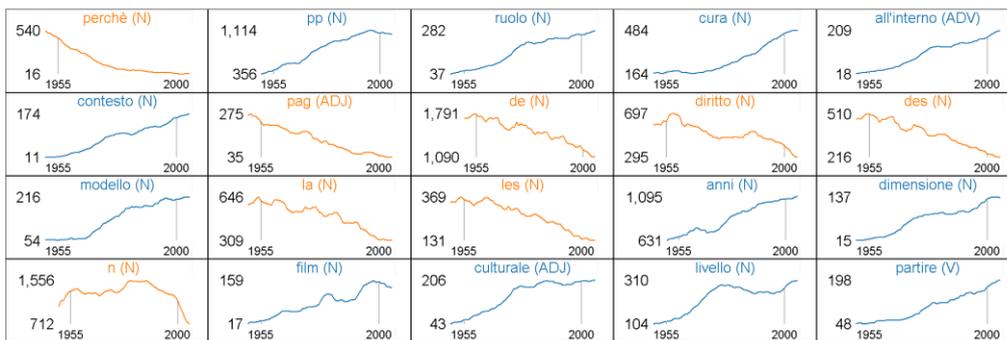


Number of volumes: 1955-1959 = 12,917 / 1995-1999 = 31,542

**Figure 12: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the German GBC**



Time series for top 20 most different

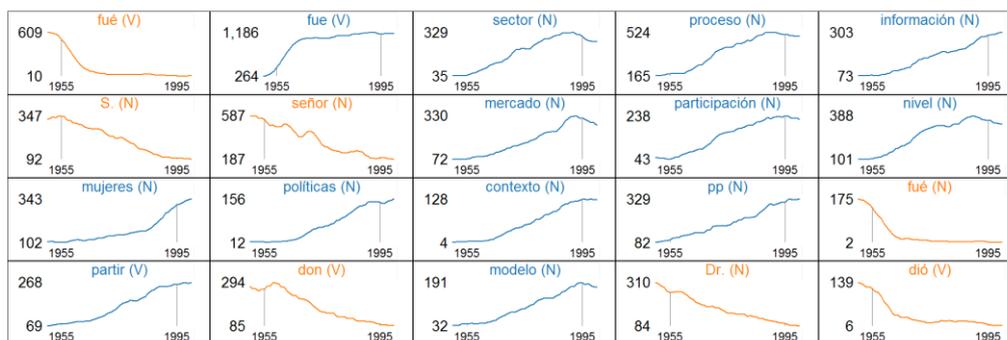


Number of volumes: 1955-1959 = 7,623 / 2000-2004 = 37,354

**Figure 13: Visualization of the differences between the time spans 1955-1959 and 2000-2004 in the Italian GBC**

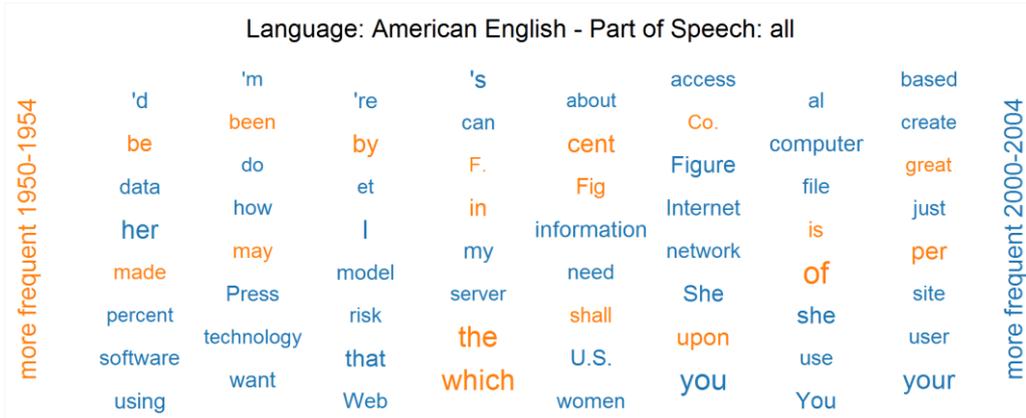


Time series for top 20 most different

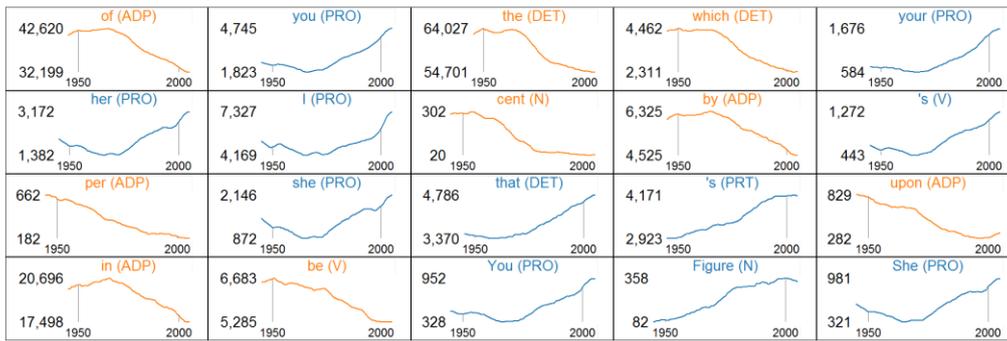


Number of volumes: 1955-1959 = 24,888 / 1995-1999 = 83,874

Figure 14: Visualization of the differences between the time spans 1955-1959 and 1995-1999 in the Spanish GBC



Time series for top 20 most different



Number of volumes: 1950-1954 = 54,633 / 2000-2004 = 422,301

Figure 15: Visualization of the differences between the time spans 1950-1954 and 2000-2004 in the British English GBC

### **3 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.**

This is a pre-copyedited, author-produced PDF of an article accepted for publication in "Digital Scholarship in the Humanities" following peer review. The version of record "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets--Reconstructing the Composition of the German Corpus in Times of WWII" is available online at:

<http://dx.doi.org/10.1093/llc/fqv037>.

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

This is a pre-copyedited, author-produced PDF of an article accepted for publication in "Digital Scholarship in the Humanities" following peer review. The version of record "Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions" is available online at:

<http://dx.doi.org/10.1093/llc/fqv030>.

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

This is a pre-copyedited, author-produced PDF of an article accepted for publication in "Corpus Linguistics and Linguistic Theory" following peer review. The version of record "Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis" is available online at: <http://dx.doi.org/10.1515/cllt-2014-0049>.

## **Eidesstattliche Versicherung**

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen

Mannheim, 13.05.2016

---

Alexander Kopenig