

Alexander Koplenig*

Against statistical significance testing in corpus linguistics

DOI 10.1515/cllt-2016-0036

Abstract: In the first volume of *Corpus Linguistics and Linguistic Theory*, Gries (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>: 285) asked whether corpus linguists should abandon null-hypothesis significance testing. In this paper, I want to revive this discussion by defending the argument that 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. As a consequence, corpus linguists should indeed abandon null-hypothesis significance testing.

Keywords: corpus linguistic methodology, statistical significance, quantitative approaches, representativeness, null-hypothesis testing

1 Introduction

In the first volume of *Corpus Linguistics and Linguistic Theory*, Gries (2005: 285) asked the following question, intended as a follow-up to Kilgarriff (2005): “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?”¹

In this paper, I want to revive this discussion by defending the argument that 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. As

¹ This is a revised and extended version of the introduction of my doctoral thesis, available at <http://hdl.handle.net/10932/00-02EC-1772-EC3F-FF01-D>, last accessed 05/20/2016. A Stata script that reproduces all results presented in this paper is available in Dataverse (doi:10.7910/DVN/ZQPWCI).

*Corresponding author: Alexander Koplenig, Department of Lexik, Institute for the German language (IDS), Mannheim, Germany, E-mail: koplenig@ids-mannheim.de

a consequence, corpus linguists should indeed abandon null-hypothesis significance testing. To this end, I want to initially describe several basic concepts that are relevant for the statistical analysis of corpus data in a non-technical but (hopefully) instructive manner (Section 2). I will then try to show why the underlying methodological assumptions are not fulfilled in corpus linguistics by surveying several propositions on the topic of representativeness that can be found in literature (Section 3). On this basis, I will briefly discuss why the use of statistical models is still highly beneficial in corpus linguistics, even without statistical significance tests, and argue for the importance of converging evidence, especially in the context of cognitive linguistics (Section 5). The paper ends with some concluding remarks in Section 5.

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. The main idea behind statistical frequentist inference is to use the distributional information from a sample of objects to estimate the characteristics of the unknown population from where the sample was taken. This is possible because under certain circumstances 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. And, as Berk and Freedman (2003: 2) put it: “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.”

As an example, we could consider a study of a certain type of flower in a certain rainforest. 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 rainforest 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; cf. Table 1). The percentage difference is $(50.0\% - 28.6\%) = 21.4$ percentage points.

As in most cases of sample-based empirical research, we are actually less interested in the specific sample but instead want to make generalizations

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

		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 %)

about, e.g. the population of this type of flower in the rainforest. 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 simple answer is that we can never be certain. For example, we could have 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 (cf. Table 2).

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

		Number of petals		
		Four	Five	
Flue	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 %)

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 a noteworthy 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. 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.

Since we know the precise size of the population in this example (something that is not usually the case), we can use stochastic calculus in order to determine the *exact* probability by counting the number of all N different 100 flower samples indexed by i with an absolute percentage difference of at least that obtained in Table 1, and divide the resulting number by the total number of possible samples with a size of 100 from a population of 1,000,000, which can be written as:

$$\begin{aligned}
 & p\left(\left|\frac{f_{fourblue}}{f_{fourblue} + f_{fourred}} - \frac{f_{fiveblue}}{f_{fiveblue} + f_{fivevered}}\right| \geq \left|\frac{50}{70} - \frac{15}{30}\right|\right) \\
 &= \sum_{i=1}^N \frac{\binom{150,000}{f_{fourblue_i}} \cdot \binom{150,000}{f_{fiveblue_i}} \cdot \binom{150,000}{f_{fourred_i}} \cdot \binom{150,000}{f_{fivevered_i}}}{\binom{1,000,000}{100}} = 0.01969
 \end{aligned}$$

This means that the probability of observing a result that is equal, or even more extreme, than the one we found in Table 1, given the fact that there is actually no relationship in the population (cf. Table 2) is very small, roughly 2%.

Figure 1 depicts the result for a simulation of *one billion* random samples of 100 flowers from the population of 1,000,000 flowers (cf. Table 2): in 196,833,77 cases of all 1,000,000,000 samples, the absolute difference in percentage is equal to or bigger than that found in Table 1; this probability ($p \approx 0.0196$) is very close to the exact solution. In roughly 72% of all samples, the percentage difference is smaller than 10 percentage points.

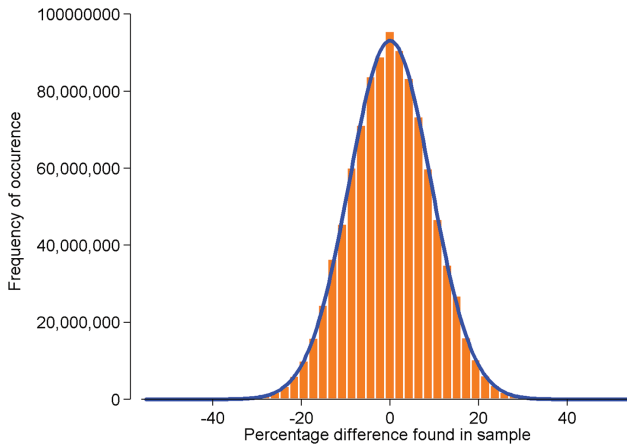


Figure 1: Histogram of percentage differences in the simulation of one billion samples.

In this context, the so-called null hypothesis (H_0) 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. This leads us to the general form of a null hypothesis statistical significance test: a result based on a random sample is called statistically significant if the probability of observing the data plus more extreme data in all potential randomly drawn samples of the same sample size is lower than some pre-selected threshold (e.g. 1% or 5%) if the null hypothesis is true and if the assumptions in the statistical test are all satisfied (Schneider 2013). Thus, the test measures the probability of observing the data given the null hypothesis is true, i.e., $p(\text{Data} | H_0)$, and not the opposite, that is, the probability that H_0 is true given the data, i.e., $p(H_0 | \text{Data})$. However, very often, this is how the test is interpreted (Cohen 1994): It does not tell us anything about how likely H_0 is (or if it is false), given the observed data, because it is calculated under the assumption that the null hypothesis is true, or more precisely, *exactly true* as in the example outlined above (Schneider 2013: 6). So, using statistical significance testing does not tell us anything about the *a priori* probability of the null hypothesis. But if we think about this (Tukey 1991: 100), it seems highly unlikely that the null hypothesis, which states that there is *exactly no* difference between, e.g. two groups, is true for any observational data. Problems associated with the idea of *exactly-no-difference null hypotheses* are discussed in Cohen (1994), Lykken (1968), Meehl (1978), Schneider (2013), or, in the context of corpus linguistics, in Kilgarriff (2005).

In general, the probability of observing the data under the assumption that the null hypothesis is true depends 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 one billion samples in the simulation led to this result. However, it also demonstrates that we can be incredibly unlucky: there is one sample in which only 6 of the 60 flowers with four petals are blue (10.00%), while 26 of the 40 flowers with five petals are blue (65.00%). 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.² 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 finding 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.³
- (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 found to be statistically significant.⁴ Some consequences of this fact

² Accordingly, in a simulation like the one presented above, 0 of one billion samples show a difference in percentage that is equal to or bigger than the difference found in Table 1.

³ Again in a simulation, almost 219 million of one billion samples show a difference of $(2/2 - 0/2) = 1$, $p = 0.219$.

⁴ It is worth noting that testing for statistical significance becomes superfluous when all members of a population are surveyed. Thus, if a corpus is treated as the population in itself, there is no need for statistical inference. In what follows (cf. Section 4), I will argue that for

that affect the analysis of corpus data will be described in Section 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. 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, than 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.

3 Corpora as representative language samples?

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).

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 3.1).
- (ii) Language data is non-random; therefore statistical approaches that assume randomness cannot be applied (cf. Section 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 3.3).

corpus linguistics “[t]reating the data as a population and discarding statistical inference might well make sense” (Berk and Freedman 2003: 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 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 3.5).

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; Kohnen 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* “is an epiphenomenon at best” (Chomsky 1986: 25).

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 “shoeboxes 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. So, I believe that Leech (2007: 3) is

right to assume that “E-language is a crucial, indispensable manifestation of I-language”.

Chomsky simply advocates the usage of the intuition of one native speaker (himself in this case) as the main “empirical” basis for the study of *I-language* (Chomsky 1986: 36–37). Wasow and Arnold (2005: 1483) 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 Stefan 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 [...] 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 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 Stefan 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 and Järvi­kivi (2007a, 2007b), Baayen

(2010), Fillmore (1992), Gilquin (2008), Gries et al. (2005), Gries et al. (2010), Kertész and Rákosi (2008), Mander 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 (cf. Arppe et al. 2010; or Ellis 2012).

3.2 Language data is non-random – statistical approaches that assume randomness cannot be applied

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. 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 $p < 0.01$) 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 demonstrates that large percentage differences are needed in order to obtain statistical significant effects in the case of small samples. It also clearly 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.

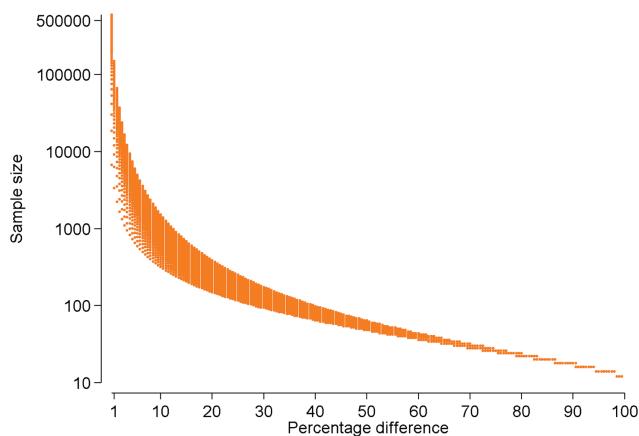


Figure 2: Estimated required sample size as a function of the percentage difference between two groups ($p < 0.01$).

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.

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 be considered as 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)

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. 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: “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” (Evert 2006: 185).

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). However, those models rely on the assumption that the texts which the corpus compiles are random samples from the textual universe approximating language as a whole. 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 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? 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 Stefan 2009: 7)?

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 is likely to affect both language reception and production to a greater extent than a non-seller. 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).

But how can we make sure that all one million people *really* listened (the whole time)? Similarly, just because a book is being sold, does not necessarily

mean that it is also being read (Ellenberg 2014). 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.

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, 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.

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: "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 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" (Biber 1993: 247), then it is completely unclear why that which is based on the traditional notion of representativeness 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” (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 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. 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 argues in reference to his library metaphor that:

In a sense, a balanced corpus is representative of the relevant sublanguage 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. [...] 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. (Evert 2006: 183; see also; Leech 2007: 4)

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 *Borgesian* 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.

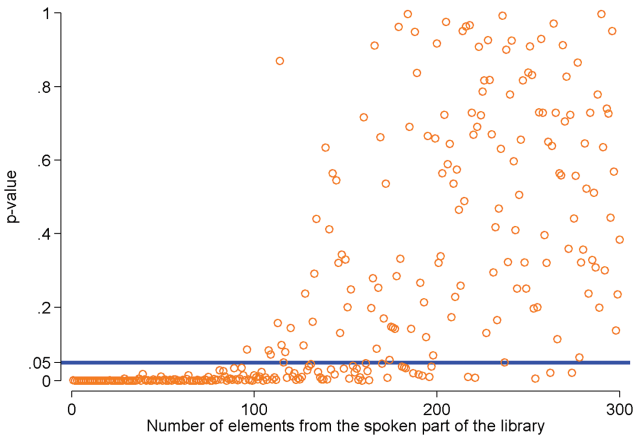
The result would lead us to believe that there is a strong relationship between the presence of X and the presence of Y in a sentence. When X is

Table 3: Result of a fictional study.

	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 %

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).

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 population. From a statistical point of view, balancing is problematic because it is subjective per definition - different researchers might

**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.

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, 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”.

This leads us directly to the next section.

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. However: “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 3.3). Those methods can and should be used to accurately describe the (linguistic) structure found in one corpus. However, the inherent problem remains: a biased sample is a biased sample. No matter how you approach it, Rieger’s claim (1979) seems right: speaking about corpora as representative samples is inappropriate. That is

why Gries's question (2005: 284) could be answered: corpus linguists should indeed abandon significance testing.⁵

In the next section, I want to discuss the consequences of this view.

4 Consequences for corpus linguistics: The importance of converging evidence and statistical analysis as a descriptive tool

To recapitulate the argument outlined in the preceding sections, probability theory helps us to quantify the amount of uncertainty that results from the data collection process. Or put differently, it gives us an idea of what is likely happen were the study to be repeated, without actually having to repeat it. However, this comes at the price of rather restrictive assumptions: (i) the researcher must be able to define the (existing) population of interest (in a non-imaginary way) from which the data are assumed to be (ii) a random sample. For corpus linguistics, both assumptions are not met.

Corpora (especially balanced ones) are samples of convenience that do not allow the extrapolation from the sample to the population, since no statistical model can compensate for this. Hence, null hypothesis significance testing is mathematically not justified in corpus linguistics.

In order to find out something 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".

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 Stefan 2009). Again, this idea points towards the importance of converging evidence (cf. Section 3.1):

one objective would be to design a method that could not only test the overall validity of a corpus model, but also the relative status of factors and interactions within that model. *If we can reproduce the ranking of factors that is suggested by a corpus-based model, that*

⁵ In the context of corpus linguistics, this is not entirely true. Using appropriate techniques, significance testing can, of course, be used if 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).

would constitute a particularly strong kind of converging evidence. Having a protocol for testing corpus results with a canon of production measures will, in the long run, earn cognitively oriented investigations of corpora a much better reputation. (Hilpert in Arppe et al. 2010: 13, my emphasis)

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.⁶

One important question remains: what does the argument outlined above imply for the statistical analysis of corpus data? Does the fact that corpora are not representative in a statistical sense also make the use of quantitative statistical models inappropriate? It does not. While a similar argumentation could be laid out for different kinds of statistical models with different types of outcome variables (linear, binary, ordinal, categorical, count, fractional, time series, etc.) or more sophisticated (mixed) models (Gries 2015: 121), let me use the standard textbook example of ordinary least square [OLS] regression, where the outcome of interest is modelled as a linear function of one or more regressors.

First and foremost, it is of utmost importance in this context to separate two steps of any empirical study that affect each other, but are nevertheless distinct methodological steps: (i) the collection of data and, after we have concluded this step, (ii) the (statistical) analysis of the collected data. Regarding (i), I have argued that, in corpus linguistics, it is not possible because of principal reasons to both well-define the population and to draw a random sample from it.

To see that, even without such a test, we can still learn something from the analysis of the data (ii), I want to re-formulate the idea of OLS regression as a concept of proportional reduction of errors for predicting an outcome [PRE]. Plot A in Figure 4 shows a scatterplot of two variables x and y . This could, for example, be a study on the relationship between two different measures characterizing the vocabulary of the inaugural speeches of all $N=8$ German chancellors. If we had to guess the value of y without knowing the value of x for any of the eight data points (black dots), our best single guess would be to use the average value of y . Plot B in Figure 4 shows that using the average value of y

⁶ 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. This eliminates the selection bias, which is the potential influence of confounding variables on an outcome of interest. Through this manipulation, it can be inferred that, random fluctuations aside, the treatment is the cause of the outcome (Diekmann 2002: 297).

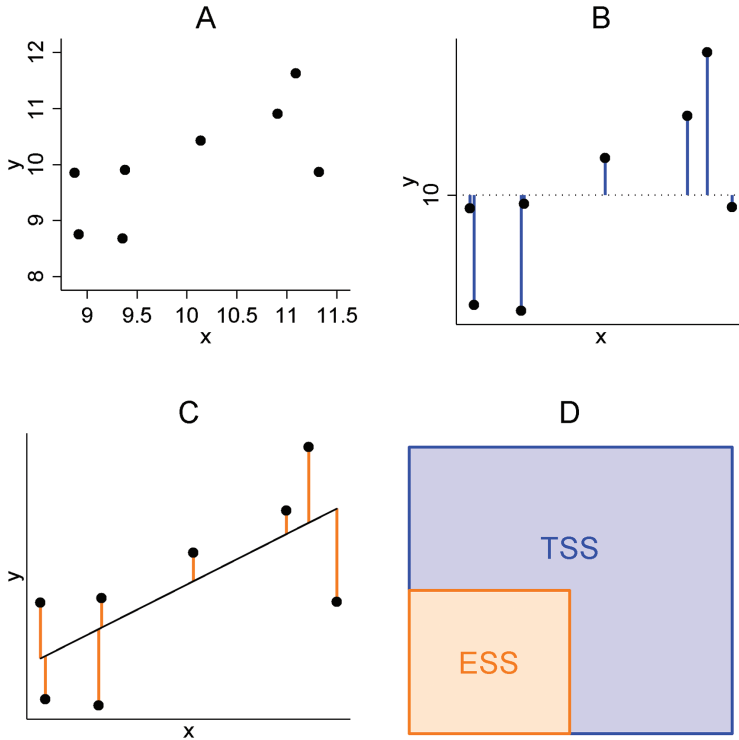


Figure 4: Result of a fictitious study. Fitting y onto x with OLS regression as a graphical method.

would sometimes be very close to the observed value of y and sometimes be quite far away from it, as indicated by the lengths of each blue line. Our prediction error can be quantified as the sum of all the squared deviations between our guess and the observed values, which is called the total sum of squares (TSS, the blue square in Plot D in Figure 4). If, however, we can use the knowledge of the value of x in order to guess the value y , and if y and x are in fact linearly related, we can improve our guess by regressing y onto x . Roughly speaking, OLS regression is just a clever way to draw a straight line on a scatter plot that “best fits” the structure of the given data set. Plot C in Figure 4 shows that if we use the x -based prediction of y to guess the value of y , our guess would still not be perfect, which is depicted by the distance between the observed values and the regression line given as $y_i = 3 + 0.7 \cdot x_i + \varepsilon_i$ for $i = 1, 2, \dots, 8$. Summing up the squares of all those deviations (ε_i , orange lines in Plot C of Figure 4) leads us to the residual sum of squares (RSS, the orange square in Plot D in Figure 4).

Now, to quantify whether knowing the value of x helps us predict the value of y can be calculated with $(TSS - RSS)/TSS$, which is in fact equal to the coefficient of determination r^2 . In the example, knowing the value of x reduces the error in predicting the value of y – on average – by roughly 49%.⁷

This is why the fact that corpora are not representative in a statistical sense does not render the use of quantitative statistical models useless, because the fitting procedure and its inherent logic have nothing to do with statistical inference. This can be easily seen by recalling that, in this example, we have collected data for all members of our population of interest (i.e. German chancellors), so there is no inference at all. However, the regression analysis still helps us to appropriately describe a potential relationship found in the data and, converging evidence again, can be used for predictions: “Forecasts about particular summary statistics, such as means or regression coefficients, can be instructive. [...] Correct forecasts would be strong evidence for the model.” (Berk and Freedman 2003: 16)

⁷ The PRE logic, in turn, can also be used to select the “best” model from a set of potential candidates: given a similar complexity of two models, the model with the better PRE value is the better choice. In this context, a so-called likelihood ratio test is quite often conducted in order to evaluate whether a more complex model (i.e. a model with additional predictors) fits the data significantly better than a less complex model. The former is called the full model, while the latter is called the constrained model. Referring to Section 2, the associated p -value does not tell us anything about whether “the constrained model is in fact true” or that “the full model fits the data significantly better than the constrained model”. It tells us something that is hard to formulate: The p -value is the probability that counts how often we would obtain – in the long run – at least the difference between the fit of the full model and the fit of the constrained model that we observed in our sample, if we were to repeat the process of randomly drawing samples of the same size from the population of interest, given both that (i) all test assumptions are met and (ii) the null hypothesis is true. I cannot provide an explanation that is fully satisfactory, but I am not sure if this testing procedure can really be justified methodologically (at all). In relation to (i), meeting test assumptions when fitting increasing complex statistical models to data dependent on the particular data set. But since fixing violations (if possible) of assumptions can also make it necessary to (slightly) modify our model specification, this means that we basically do not always obtain the same full and constrained models for our (hypothetical) samples for which we then (hypothetically) compare goodness of fits. Even for a OLS regression, the assumption that permits a test for statistical significance concerns the distribution of errors, something which “is not empirically identifiable outside the model” (Berk and Freedman 2003: 9). In relation to (ii), we have to keep in mind that including additional predictors in a model will almost always (at least slightly) improve its fit. Thus, the null hypothesis will also almost always not be true, as it seems to claim precisely this; that there is *exactly* no difference between the fit of the full model and the constrained model. If, however, the null hypothesis is almost always not true, then we cannot – almost always – assume it to be true in a test for statistical significance.

In this case, it would be, for example, particularly strong evidence for our model, if our model-based regression coefficient could help us to predict the vocabulary of the inaugural speeches of prospective German chancellors.

Thus, statistical models can be used as a descriptive tool with great explanatory power in order to squeeze out as much information as possible from the data without drawing inferences to a population and without using (arbitrary) *p*-values to make claims about the reproducibility of scientific findings.

5 Concluding remarks

To see that the discussion about the problems associated with *p*-values is not something specific to corpus linguistics, take for instance, one of *Nature's* most viewed articles,⁸ where Nuzzo argues that: “P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume” (Nuzzo 2014: 150).

Or, take the statement from the ASA, the world’s largest professional association of statisticians published in 2016: “The widespread use of ‘statistical significance’ (generally interpreted as “ $p \leq 0.05$ ”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process” (Wasserstein and Lazar 2016: 9).

In this paper, I have argued that corpus linguists should abandon statistical significance testing, because its random-sampling assumptions are not applicable. Instead of using statistical models to draw inferences about the studied language, I suggest that corpus linguistics should start to produce a different kind of standard by combining corpus data with other empirical linguistic information in order to strengthen its methodological basis and – in the context of cognitive linguistics – in order to bring us closer to “the ultimate aim of cognitive linguists [... :] to come up with linguistic models that actually claim to reflect (what we believe to know about) the way our minds work.” (Schmid 2010: 102).

In addition, I have tried to demonstrate that even without testing for statistical significance, corpus linguistics can still greatly benefit from the use of quantitative statistical methods as a tool to describe the structure found in corpora.

I hope that this paper will trigger a fresh discussion on the issue of null-hypothesis significance testing in corpus linguistics.

⁸ <https://www.altmetric.com/details/2115792> (last accessed 05/20/16).

Acknowledgments: I would like to thank Arnulf Deppermann, Ludwig M. Eichinger, Stefan Engelberg, Martin Hilpert, Matthias Kohring, Peter Meyer, Sarah Signer, Carolin Müller-Spitzer and Sascha Wolfer for their input and feedback. All remaining errors are mine.

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 & Järvi­kivi Juhani 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.
- Arppe, Antti & Järvi­kivi Juhani 2007b. Every method counts – Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.
- 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.
- 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 edn. New York: Aldine de Gruyter. <http://www.stat.berkeley.edu/~census/berk2.pdf> (accessed 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 (accessed 30 March 2015).
- Brezina, Vaclav & Miriam Meyerhoff. 2014. Significant or random? A critical review of socio-linguistic 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/> (accessed 21 October 2014).
- Chomsky, Noam. 1986. Knowledge of language: Its nature, origin, and use. In *(Convergence)*. New York: Praeger.
- Cohen, Jacob. 1994. The earth is round ($p < 0.05$). *American Psychologist* 49(12). 997–1003. doi:10.1037/0003-066X.49.12.997.
- 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> (accessed 10 June, 2015).

- Diekmann, Andreas. 2002. *Empirische sozialforschung: Grundlagen, methoden, anwendungen*. 8th edn. 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 already (and how)? In Jost Gippert & Ralf Gehrke (eds.), *Historical corpora: Challenges and perspectives* (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5), 13–33. 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> (accessed 11 June 2015).
- Ellis, Nick C. 2012. What can we count in language, and what counts in language acquisition, cognition, and use? In *Frequency effects in language learning and processing*. Berlin & Boston: de Gruyter (accessed 19 May 2016).
- Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–190. (accessed 19 March 2014).
- Fillmore, Charles J. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics." In Jan Svartvik (ed.), *Directions in Corpus Linguistics*, 35–60. Berlin: de Gruyter.
- 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 & Th. Gries Stefan. 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> (accessed 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 & Schönefeld. Döris 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 & Schönefeld. Döris 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.
- Harald, Baayen, R. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5. 436–461.
- Hunston, Susan. 2010. *Corpora in applied linguistics* 7. print. (The Cambridge Applied Linguistics Series). Cambridge: Cambridge University Press.
- Jann, Ben. 2005. *Einführung in die Statistik*. München & Wien: Oldenbourg.
- 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. 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>.

- Köhler, Reinhard. 2005. Korpuslinguistik – zu wissenschaftstheoretischer Grundlagen und methodologischen Perspektiven. *LDV Forum* 20(2). 1–16.
- Kohonen, 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/kohonen> (accessed 5 October 2014).
- 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. Amsterdam: Rodopi.
- Lijffijt, Jeffrey, Terttu 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> (accessed 22 April 2015).
- Lykken, David T. 1968. Statistical significance in psychological research. *Psychological Bulletin* 70(3, Pt.1). 151–159. doi:10.1037/h0026141.
- 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 (accessed 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.
- Meehl, Paul E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Counseling and Clinical Psychology* 46. 806–834. doi:10.1016/j.appsy.2004.02.001.
- Nuzzo, Regina. 2014. Scientific method: Statistical errors. *Nature* 506(7487). 150–152. doi:10.1038/506150a.
- Oakes, Michael P. 1998. *Statistics for corpus linguistics* (Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University 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* (Monographien Linguistik Und Kommunikationswissenschaft 39), 52–70. Königstein im Taunus: Scriptor. <http://www.uni-trier.de/fileadmin/fb2/LDV/Rieger/Publikationen/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.
- Schneider, Jesper W. 2013. Caveats for using statistical significance tests in research assessments. *CoRR* abs/1112.2516.

- 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* (Human Cognitive Processing v. 33), 1–31. Amsterdam & Philadelphia: John Benjamins Pub. Co.
- Schütze, Carson T. 1996. *The empirical base of linguistics*. Chicago: The University of Chicago Press.
- Trochim, William. 2006. Design. *Research Methods Knowledge Base*. <http://www.socialresearchmethods.net/kb/design.php> (accessed 14 September 2011).
- Tukey, John W. 1991. The philosophy of multiple comparisons. *Statistical Science* 6(1). 100–116. doi:10.1214/ss/1177011945.
- 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.
- Wasserstein, Ronald L. & Nicole A. Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. doi:10.1080/00031305.2016.1154108.
- Wiechmann, Daniel. 2008. On the computation of collocation strength. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.