Exploring variability within and between corpora: some methodological considerations[1]

Stefan Th. Gries[2]

Abstract

The results usually reported in corpus-linguistic studies are quantitative: frequencies, percentages, model parameters, *etc*. However, given that no corpora are alike, and that sometimes different results are reported for very similar corpora (or even the same corpus), three central issues are: (i) how to identify and quantify the degree of variation coming with one's results; (ii) how to investigate the source of the observed variation in corpora; and, (iii) how homogeneous one's corpus is with respect to a particular phenomenon.

In this paper, I shall present a methodology that addresses these issues, providing data from ICE-GB on the frequency of the English present perfect, the alternation of transitive phrasal verbs and the semantics of the English ditransitive. Specifically, I will show how applying resampling methods and exploratory data analysis to corpus data allows for, (i) providing interval estimates for one's findings that show how superficially different results may reflect similar underlying tendencies; (ii) determining communicative dimensions underlying variation in a bottom-up fashion (similar to work by Biber, but based on just the phenomenon one is interested in); and, (iii) quantifying the homogeneity of the corpus with respect to the phenomena one is actually interested in (rather than by the standard approach of using word frequencies).

> For every parameter we estimate from data, we need to establish an *unreliability estimate*. We use this to judge the uncertainty associated with any inferences we may want to make about our point estimate, and to establish a confidence interval for the true value of the parameter. Up to now, we have used parametric measures like standard errors that are based on the assumption of normality of errors

[1] I thank three anonymous reviewers as well as audiences at the workshop 'Corpus statistics: Objectives, Methods, Problems' at the University of Leipzig (September 2005), the second workshop of the Construction Grammar Network Germany at the University of Jena (October 2005), and at ICAME 27 in Helsinki (May 2006) for their feedback on versions of this paper. The usual disclaimers apply.

[2] *of* Department of Linguistics, University of California, Santa Barbara, CA 93106–3100, USA
Correspondence to: Stefan Th. Gries, *e-mail*: stgries@linguistics.ucsb.edu

> […]. If the assumption of normality is wrong, then our unreliability estimates will also be wrong, but it is hard to know how wrong they will be, using standard analytical methods. An alternative way of establishing unreliability estimates is to *resample* our data […]
>
> (Crawley, 2002: 195; emphasis as original)

1. Introduction

One of the most important concepts within corpus linguistics is variability. Variability is a key issue on several levels, simultaneously. First, variability is always of prime importance when reporting one's results: without an indication of the variability found in one's data, the interpretation of, say, aggregated frequencies/percentages or measures of the central tendency of a single study is usually quite difficult, and the comparison of results between different studies is seriously impaired.

Secondly, variability is an essential issue because corpora are inherently variable internally. This point touches on the issue of corpus homogeneity and is concerned with the fact that most phenomena of interest will yield different results when investigated in different parts of a corpus. Thirdly, corpora are also variable externally. I am, of course, referring to the fact that results concerning the same phenomenon will differ between different corpora, which in turn makes it difficult to generalise from results obtained from one corpus (part) to other corpus parts or even different corpora. There is some work in this area, some of which I shall mention in more detail below, but this complex of problems appears not to have received the amount of attention it deserves within our discipline. This is all the more perplexing since these problems not only show up in the day-to-day reading and writing of corpus-based analyses, but they also surface regularly in public: probably every corpus linguist gets to hear the following kinds of questions after talks: 'Isn't this just true in oral/written/… language?' and 'Isn't this just because you only looked at genre *XYZ*?' *etc*. These questions raise the often legitimate point about whether the data analysed are so variable that, perhaps, a different register would yield somewhat different or even opposite results, and may even outweigh those that were reported. Note, however, that without any empirical evidence such questions are in two ways just guesses. First, there are several levels of hierarchical organisation or granularity at which variability might be located: modes, registers, sub-registers (see below) or even lexically-defined levels. Secondly, even within one level of hierarchical granularity there are usually more than two levels between which differences may exist. Thus, for instance, even if differences are located at the level of the register, this need not mean that all registers are (equally) different form each other; in the language of ANOVAs, even if a factor is significant, this does not mean that all its levels make a difference.

Let me give a brief example of what I have in mind. Schlüter (2005) compared different corpus-based studies on the overall frequency of the present perfect in English. Table 1 provides an overview of his findings; to his values I have added (in brackets) *z*-scores that provide each frequency's deviation of the mean of all frequencies in the number of standard deviations.[3]

Author (*year*): *corpus* (*parts*)	*Present perfect in 1,000 words*	*Author* (*year*): *corpus* (*parts*)	*Present perfect in 1,000 words*
Dubois (1972): Brown E	4.5 (0.31)	Dubois (1972): Brown F	1.2 (-1.28)
Elsness (1997): Brown E	3.1 (-0.36)	Elsness (1997): Brown F	0.7 (-1.52)
Schlüter (2002): Brown E	4.1 (0.12)	Schlüter (2002): Brown F	1.3 (-1.23)
Elsness (1997): LOB E	3.5 (-0.17)	Elsness (1997): LOB F	1.1 (-1.33)
Schlüter (2002): LOB E	4.9 (0.5)	Schlüter (2002): LOB F	2.1 (-0.84)
Biber *et al*. (1999): conversation	5.9 (0.99)	Herzlík (1976): expos. prose	3.5 (-0.17)
Mindt (2000): conversation	6 (1.04)	Mindt (2000): expos. prose	5 (0.55)
Mindt (2000): fiction	2 (-0.89)	Schlüter (2002): CEC	5.9 (0.99)
Herzlík (1976): three novels	3.6 (-0.12)	Biber *et al*. (1999): news	6.1 (1.08)
Herzlík (1976): one drama	9 (2.48)	Biber *et al*. (1999): academic	4 (0.07)
Biber *et al*. (1999): fiction	3.4 (-0.22)		

Table 1: Frequencies of present perfects in previous studies (from Schlüter, 2005)

There are several things to notice here. First, there is quite some variability within the data. The mean of all relative frequencies is 3.85, but

[3] I am grateful to Norbert Schlüter for discussion and making his data available to me.

the values range from 0.7 to 9.[4] Specifically, the *z*-scores span a range of exactly four standard deviations: from -1.52 for Elsness's count in Brown F to 2.48 for Herzlík's count for a drama. Secondly, there is a large variation within some registers. For example, while the frequencies for conversation are nearly identical, the frequency for drama, which is in these studies often included in the conversation data (Schlüter, personal communication), is about 50 percent higher. Also, the frequencies for expository prose differ considerably and only the fiction data yield rather homogeneous results. Thirdly, it is perturbing to find that the results sometimes display large discrepancies for data from the same corpus. For example, the larger proportion for LOB F is about twice as high as the smaller one, and the same holds for the largest and the smallest proportion of Brown F. The results for Brown E are somewhat less extreme, but even here the largest figure is about 50 percent higher than the smallest. All of these observations cast serious doubt on the results and, on that basis, Schlüter raises the (legitimate) question of the reliability of corpus data.

However, there are also three problems inherent in the data. The most basic of these has been mentioned above: it is well-known to any statistician that reporting frequencies or means without adding an index of dispersion of the frequencies (or means) is misleading since it is unclear how well the single reported frequency (or mean) represents the data it is supposed to summarise.

Another point is that the choice of the ratio – present perfects per 1,000 words – is not particularly useful since it alone may be responsible for much of the observed variability. If the corpora do not only differ with respect to their proportion of present perfects but also with respect to the number of verbs they contain, this will also be reflected in the present perfect ratio of Table 1, which is why the frequency of present perfects is better expressed as the number of present perfects per 100 or 1,000 verbs. This assumption, which also threatens other studies' validity (such as Berglund, 1997), is indeed proven right by Schlüter (personal communication) and renders any interpretation of these results doubtful.

The third problem is, however, more fundamental in nature and arises even if the first two problems have been resolved. It concerns the main issue of this paper, namely the question of which degree of variability one would expect anyway both *within* every individual study and between the different studies cited by Schlüter. For example, if the within-study variability associated with each mean value presented in Table 1 was large, then the differences *between* the results for the different corpora would probably not be too important. Thus, the frequency of present perfect would have to be considered very variable or volatile even within a single data set. If, on the other hand, each mean from Table 1 came with a small

[4] I am glossing over the fact that one would have to compute a weighted mean here, figuring the different corpus sizes into the computation of the overall mean. This point does not bear on the subsequent discussion.

degree of variability, then the differences between different files or registers would be more likely to indicate linguistically-relevant differences, which in turn would mean that the frequency of the present perfect is quite variable, but could be explained with respect to different registers/genres or any other parameter of interest.

This paper addresses this complex of interrelated questions. I hope to stimulate a more comprehensive discussion of such issues by proposing a family of methods which can be used to answer the following questions: if one performs a corpus-linguistic quantitative analysis of phenomenon *X* in a corpus using the statistical parameter *P*,

(i) from a descriptive perspective, which degree of variability of *P* was observed in one's data set and how do we quantify it? And, how do the present results concerning *P* compare to those of other studies? These questions inevitably lead to the next:
(ii) how homogeneous is the corpus that was used for the study of phenomenon *X*? And,
(iii) from an exploratory, bottom-up perspective, how can one identify (some of) the (most) relevant sources of the observed variability of *P*? If we rephrase that from a hypothesis-testing, top-down perspective: are the variables A, B, C, responsible for a significant proportion of *P*'s variability?

The paper is structured as follows. My suggestions below do not arise out of a vacuum and at least some of them can be viewed as having evolved from previous work in this domain. In Section 2, I will, therefore, discuss some previous work that was concerned with similar issues (but mostly with very different techniques). Sections 3 and 4 introduce two case studies which will exemplify various aspects of the methods to quantify and explore variability in corpora. While these two sections will also touch on the issue of corpus homogeneity, Section 5 is specifically devoted to this topic and will provide a more sophisticated technique for measuring corpus homogeneity. Section 6 will summarise and conclude.

2. Previous approaches

In spite of the importance and omnipresence of the issue of variability in virtually all corpus-linguistic results, there are many studies which report quantitative results, such as means or frequencies, without any kind of statistical testing and/or assessment, and subsequent interpretation of the variability of the data they have summarised (see Berglund, 1997; Boas, 2003; Egan, 2002; Facchinetti, 2001; Hunston and Francis, 2000; Johansson, 2001; Kennedy, 1991; Mukherjee, 2003; and many more). It is even worse to find that there is not much work that systematically explores the issues of variability within corpora (i.e., corpus homogeneity) and between corpora, 'to this date we do not know of a satisfying definition of corpus homogeneity' (Denoual, 2006: 5). It is important to bear in mind from the outset that variability between and within corpora requires one to make two decisions: first, a decision concerning the parameter whose variability will be measured, (which has mostly been in the form of word frequencies); and secondly, a decision concerning the level of granularity at which the corpus will be investigated, (which has mostly been the level of the file or a register). In this section, I will mention briefly some of the studies that have been devoted to these topics to set the stage for my own stance on these issues.

There are a few studies that have been concerned with variability *between* corpora. One of the earliest is Hofland and Johansson (1982) who study word frequencies in British and American English, and use the Brown and LOB corpora to identify words that are more typical of one variety of English than the other. They use the chi-square test as well as Yule's difference coefficient. Johansson and Hofland (1989) improve on this work by factoring in word-class information. Leech and Fallon (1992) adopt a very similar approach, identifying words that are significantly more frequent in British or American English to discuss cultural differences. Rayson and Garside (2000) use the log-likelihood statistic to compare a target corpus of air traffic control communication against a reference corpus (a part of the spoken component of the British National Corpus), a technique that is now also available as part of Mike Scott's corpus suite WordSmith Tools. Oakes (2003) does very much the same as the early studies, but he uses two other corpora, namely FROWN and FLOB, and he very briefly tests another method: high ratio pairs. (See Rayson (2003) for a detailed overview of this kind of lexically-based approach.) Finally, Denoual (2006) proposes a method to quantify corpus similarity based on the cross-entropy of statistical *n*-gram character models.

As to studies more directly concerned with variability *within* corpora, i.e., corpus homogeneity, there are a few works to be mentioned briefly. Rayson, Leech and Hodges (1997) use the conversational part of the British National Corpus (BNC) and conduct chi-square tests to determine word-frequency differences between female vs. male speakers, as well as between speakers of different ages and social groups. However, perhaps the most frequently-cited study in this subject area is Kilgarriff (2001), whose approach is based on:

> a set of "Known-Similarity Corpora" (KSC). A KSC-set is built as follows: two reasonably distinct text types, A and B, are taken. Corpus 1 comprises 100% A; Corpus 2, 90% A and 10% B; Corpus 3, 80% A and 20% B; and so on.
>
> (Kilgarriff, 2001: 121)

Given the set of corpus similarity statements inherent in this data set, Kilgarriff compared several similarity measures and found that, with some caveats, Spearman's ρ outperformed chi-square, which in turn outperformed the information-theoretic notion of perplexity. In that paper, and also in a very similar follow-up (Kilgarriff, 2005), he uses data from a comparison of two corpus parts of the BNC to show that the kind of chi-square-based word-frequency comparisons is bound to yield an extremely high number of false positives, i.e., significant results which are due to the large sample size but do not matter much. See Rose, Haddock and Tucker (1997) for similar work, which also briefly explores a bottom-up approach based on word frequency lists; De Roeck, Sarkar and Garthwaite (2004) for an extension of Kilgarriff's work experimenting with three different parameters underlying the partitioning of the data sets (documents, document halves and various chunk sizes); and Sahlgren and Karlgren (2005) for a similar investigation using a density measure.

Gries (2005b), in reply to Kilgarriff (2005), addresses corpus homogeneity more indirectly, concentrating instead on some general methodological issues of word-frequency hypothesis testing. First, Gries argues in favour of using effect sizes as the most relevant measure. Secondly, since Kilgarriff's word frequency study involves just a single comparison and does not explicitly invoke effect sizes, Gries reports the results of a 'simulation study' involving forty-five pairwise word-frequency lists of the ten largest files of the BNC, comparing regular chi-square tests, chi-square tests with a *post hoc*-correction and three measures of effect size. Gries finds that while regular chi-square tests do yield more significant results than one would expect, *post hoc*-corrected *p*-values approximate the null hypothesis ideal of 5 percent quite well. Also, the effect sizes obtained for the word frequency tests show quite clearly that most of the significant effects were practically irrelevant, concluding that the overall outlook on such lexically-based approaches may not be as pessimistic as suggested by Kilgarriff.

By way of an interim summary, the vast majority of these previous approaches are based on chi-square word-frequency testing involving Brown, LOB, Frown, FLOB and the BNC. Unfortunately, this fact already highlights some problematic aspects of these works. One is that operationalising homogeneity as deviations of observed word frequencies from expected word frequencies presupposes that: (i) word frequencies are independent, and (ii) expected word frequencies can be derived from a maximum likelihood estimate. But it is clear that this is not so (see Church, 2000). Thus, the reliance on the kind of chi-square testing that is so prominent in this area is unfortunate and an approach utilising effect sizes may ultimately be more desirable.

Another problematic aspect is that most of these studies are based on word frequencies. This methodological choice is easy to understand given that word frequencies are comparatively easy to recover from corpora. However, for two reasons, this choice also seriously limits the range of applicability of these approaches. First, an approach to corpus homogeneity based on word frequency is much more likely to produce biased results when applied to corpora containing text samples focusing on a particular topic.[5] Secondly, interesting as corpus variability/homogeneity results may be on a lexical level, they have little or nothing of interest to offer a linguist who is primarily interested in grammatical or other phenomena. Such researchers would gain nothing by investigating some grammatical phenomenon but basing their assessments of variability and/or corpus homogeneity on word frequencies. There are a few studies, however, which go beyond the simplest word-frequency approach.

One study that is quite similar in spirit to what will be discussed shortly, although it had a rather different focus, is Biber (1990). He is concerned with issues of corpus compilation such as the number and the length of texts needed to identify and represent the linguistic characteristics of text types. At a time when most studies were still restricting themselves to comparing word frequencies across corpora or corpus parts, Biber split up corpus samples into parts to determine the degree to which grammatical features exhibit similar frequency distributions and factorial structures. Even though Biber's study has had quite an impact on corpus compilation, its implications are far reaching and have unfortunately not received the recognition they deserve. In this connection, it is interesting to note, however, that there is some first evidence that sometimes even word-frequency-based approaches can reveal the kind of register differences illustrated by Biber's sophisticated analysis (see Gries, 2005b: Section 3.2; Xiao and McEnery, 2005).

The work by Sekine (1997) is also similar to what will be proposed below. Sekine investigates genre differences in the Brown corpus on the basis of cross-entropy values from partial (depth 1) syntactic trees. Another study which is methodologically similar in spirit to what follows below, but is concerned with something very different, is Evert and Krenn (2005), who use a random sample evaluation approach to test which association measure is best suited to the automatic identification of German PP-verb constructions. Thus, their paper does not involve variability between and within corpora in exactly the sense used here, but I will make use of a similar randomisation method below. Finally, Gries, Hampe and Schönefeld (forthcoming) use a Monte-Carlo-like simulation technique to compare the degree of association of particular verbs to the so-called *as*-predicative (e.g., 'He regards that as a stupid idea') in the BNC Sampler as opposed to in the British Component of the International Corpus of English (ICE-GB, Release 1). They test whether the number of overlapping verbs, i.e., verbs attracted to this construction in both corpora, could be obtained by chance by comparing the observed overlap to the expected one, namely that obtained in 100,000 samples without replacement (where the sampling was weighted by verb frequency). As a result, they obtain an index of the degree to which a construction differs in two corpora that is independent of all words' frequencies, *etc*.

[5] I thank an anonymous reviewer for raising this point.

In this paper, many of the methods outlined below are inspired by the epigraph at the beginning of the paper and, thus, based on methods from exploratory data analysis using summary statistics, data reduction and structuring methods, as well as graphical methods. The focus will be on effect sizes rather than on significance testing. In addition, and as just mentioned, I will also make some proposals that involve resampling approaches. The notion of resampling refers to:

> a variety of methods for computing summary statistics using subsets of available data (jackknife), drawing randomly with replacement from a set of data points (bootstrapping), or switching labels on data points when performing significance tests (permutation test, […]).
> (Wikipedia contributors, 2006)

Of these methods, I will use the second and third: bootstrapping and permutations. One key aspect of the implementation of such methods here is that it will not only involve sampling and permuting arbitrarily-defined corpus parts (e.g., files), but also parts manifested at levels of granularity which a researcher suspects are important. I would also like to stress from the outset that the methods proposed below require neither much data beyond what most corpus-linguistic methods already provide nor a huge set of software applications: most of the methods can be performed with what is available from simple concordance lines, and all of the data, analyses and graphics found below were obtained using just one piece of software, namely R (R Development Core Team, 2005), an open source programming language and environment for statistical computing.

Linguistically, the emphasis will be on handling variability and corpus homogeneity *solely on the basis of the parameter*(*s*) *an analyst is really interested in* rather than on the so far predominant word frequencies or on any other supposedly universally applicable parameter that happens to be available (e.g., Sekine's depth-1 trees, Denoual's character *n*-grams *etc*.). This is not at all to downplay these other approaches. My point is just that a method that is specifically geared to what a linguist with a particular theoretical interest – such as present perfects – may need is more likely to yield interesting results for that particular linguist than any other method. Thus, if one investigates present perfects, why should one care about word frequencies or character *n*-grams?

3. Case study 1: the frequency of present perfects in English

This section will take up Schlüter's results mentioned in the introduction of this paper. I will propose a variety of progressively more complex and comprehensive ways of quantifying and exploring the variability of corpus results on the basis of frequencies of the present perfect in the ICE-GB.

Since the most natural way to quantify the variability of the data consists of looking at how much individual corpus parts differ from each other (recall that most studies used genres or files) and since Schlüter's study involves register/genre distinctions, I will also base many of the proposals that follow on that level of granularity. However, as will become apparent, nothing hinges on choosing this as the relevant level of granularity. In what follows, I will propose two methods: one involves a very simple and coarse understanding of 'individual corpus parts' (Section 3.1), and the other is more complex but provides results at a higher level of granularity or precision (Section 3.2).

3.1 Simple approaches to describing variability

The first method involves the use of simple descriptive statistical techniques. I first generated a table that lists every file from the ICE-GB together with information concerning its medium, its register and what I will refer to as sub-register. The distinctions established on these three levels are presented in Table 2; the finest level of distinction possible forms thirteen sub-registers.

Secondly, I computed for each file the percentage of all verb forms in the present perfect out of all verb forms, and assembled this information in a table listing all files and their register properties. The result is presented in shortened form in Table 3.

Medium	*Register*	*Sub-register*
Spoken	Dialogue	Private, public
	Monologue	Scripted, unscripted
	Mix	Broadcast
Written	Printed	Academic, creative, instructional, nonacademic, persuasive, reportage
	Non-printed	Letters, non-professional

Table 2: Medium, register and sub-register distinctions in the ICE-GB

Finally, I computed the average frequency of present perfects of all files for each sub-register. In addition, and to be able to compare the results to those of Schlüter in at least some way, I also computed the corresponding *z*-scores, which are presented in Table 4.

Medium	*Register*	*Sub-register*	*File*	*Present perfects*	*Verbs*
spoken	dialog	private	S1A-001	…	…
…	…	…	…	…	…
spoken	monolog	scripted	S2B-035	…	…
…	…	…	…	…	…
written	print	reportage	W2C-001	…	…
…	…	…	…	…	…

Table 3: File, medium, register and sub-register distinctions in the ICE-GB

Sub-register and numbers of files per sub-register		*Frequency*	*z-score*
{written printed creative}	20	0.0123	-1.46
{written printed instructional}	20	0.0164	-1.16
{written printed nonacademic}	40	0.0172	-1.09
{written printed academic}	40	0.0226	-0.69
{written nonprinted nonprofessional}	20	0.0291	-0.20
{spoken dialogue private}	100	0.0301	-0.13
{spoken dialogue public}	80	0.0331	0.10
{spoken monologue unscripted}	70	0.0343	0.19
{written printed reportage}	20	0.0349	0.23
{written nonprinted letters}	30	0.0375	0.43
{spoken monologue scripted}	30	0.0377	0.44
{written printed persuasive}	10	0.0461	1.07
{spoken mix broadcast}	20	0.0623	2.28

Table 4: Relative frequencies of present perfects for all sub-registers in the ICE-GB (sorted in ascending order)

Several things are worth noting. First, the range of *z*-scores provides us with an index of the variability of the data. This is much more informative than a single standard deviation, or confidence interval of the overall percentage of the entire corpus, could be.[6] Secondly, once we have such results, we can now also observe that the findings obtained from just a single corpus do not differ much from what was obtained in the many different studies reviewed by Schlüter, which is of course due to the fact that the compilers of the ICE-GB aimed to represent a broad variety of registers and genres: the *z*-scores span a range of 3.74 standard deviations (corresponding roughly to the four standard deviations found in Schlüter's data), namely from -1.46 for {written printed creative} to 2.28 for {spoken mix broadcast}, and even the minimum and maximum *z*-scores are quite similar: -1.46 and 2.28 for the sub-registers of the ICE-GB are very close to the -1.52 and 2.48 found in Schlüter's data. All of this shows that while the variable findings *for one and the same corpus* are certainly problematic and point to disparities of data collection and/or evaluation, the overall amount of variability reported by Schlüter *for different corpora* is far from unusual and can be expected even from sampling different parts of a single corpus intended to represent British English of the 1990s.

Splitting the corpus into sub-registers to compute percentages and *z*-scores not only allows us to see how different the values are, it also allows us to determine the sub-registers which are closest to, or furthest from, the percentage for the overall corpus. The percentage for the overall corpus is arrived at by dividing the overall frequency of present perfects by the overall number of verbs, which yields 0.0308. The sub-register coming closest to the value for the whole corpus is {spoken dialog private} with a percentage of 0.0301. This is an interesting finding for two reasons. First, this sub-register is often regarded as the most basic form of human language and it appears to be a striking coincidence that it is just this sub-register that is closest to the overall average proportion of present perfects. Secondly, this sub-register is often considered to be one of the most inherently variable, which makes it all the more perplexing that it is so close to the corpus's overall mean.

[6] An easy way to solve this problem may seem to be to provide just a measure of dispersion together with means or frequencies (e.g., a standard error, a standard deviation, a confidence interval, …), as was also suggested by an anonymous reviewer. However, I disagree on two grounds. Let us assume that every all present perfects are coded as 1 and all other verb forms are coded as 0. First, I specifically argue against the use of a *single* standard deviation for the complete corpus (0.1728 in this case) because while this single figure does of course provide an overall index of variability, unlike separate *z*-scores as those in Table 1 or Table 4, it does not allow further exploration as to where below-average and above-average frequencies occur, which groups of sub-registers seem to exist *etc*. Second, even if one used *separate* standard deviations for the different corpora of Table 1 (or the different corpus parts of Table 4) problems can still arise because one cannot straightforwardly compare the standard deviations to each other: the standard deviations are associated with different means (and are highly correlated with them) so the measure one would need is the variation coefficient, each standard deviation divided by its mean. Thus, I submit that the range of *z*-scores is superior to both one standard deviation or one for each sub-register. Be that as it may, I will outline more sophisticated approaches below.

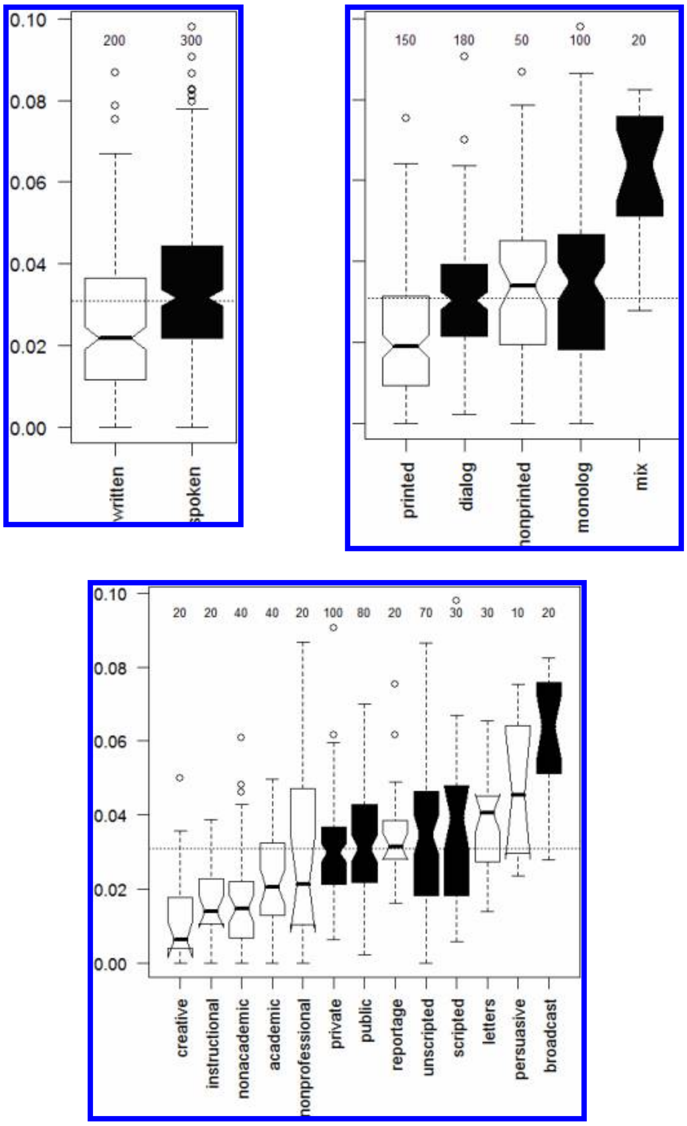


Figure 1: Boxplots of mean frequencies of present perfects in the two modes, all registers and all sub-registers of the ICE-GB[7]

[7] A boxplot is a plot which summarises the central tendency, the dispersion, and the overall distribution of a set of values. In the present kind of boxplots the thick horizontal line within a box indicates the median of the data points; the lower and upper limit of the box indicates the lower and upper hinge (the medians of the data points below and above the median); the whiskers extend to the most extreme data point which is no more than 1.5 times the length of the box away from the box; as in much other software, the notches extend to $\pm 1.58 \cdot {}^{IQR}/_{sqrt\ n}$ to provide an approximation of a 95 percent confidence interval for the difference in two medians. Thus, non-overlapping notches provide strong evidence that the medians differ significantly from each other (cf. R's help page, s.v. *boxplot*).

While this approach is already an improvement on reporting the percentages, it still has shortcomings that need to be addressed. One is that we have so far neglected the variability within the sub-registers. The other is that we are currently conducting only basic individual comparisons of frequencies at the level of the thirteen sub-registers, although it is unclear whether this is actually the level of granularity where the most relevant source of variability can be observed; cf. Section 1. It is worth remembering that this of course also applies to the many studies mentioned in Section 2, which also used sub-registers, domains, files, *etc*. A minimal extension, though, allows me to address both of these issues. My proposal is to summarise the frequencies of present perfects by generating boxplots for every level of granularity suspected to be important. Figure 1 is an example of where the two modes, the five registers and the thirteen sub-registers can all be compared at the same time to determine immediately the parameters that are responsible for most of the variability in the data. The figures above the upper whiskers denote the number of corpus files summarised in the box below the figure; black and transparent boxes indicate spoken and written data respectively; the horizontal dotted lines indicate the overall mean of 0.0308.

The results are somewhat interesting because they indicate what is already possible in a case study that is small and not statistically sophisticated. In the first plot, we find that the percentage of present perfects in speaking is about 1 percent higher than in writing, which is a significant difference. The second plot, however, shows that this general statement should be qualified: {spoken dialog} and {spoken monolog} do not differ significantly from {written nonprinted}, and both are also fairly close to {written printed}. However, {spoken mixed broadcast} is an outlier with a much higher percentage than the other registers of the corpus, suggesting that a more detailed analysis of this corpus part might be revealing.

The final plot suggests that the register {written printed} consists of two parts. One is relatively homogeneous and comprises {written printed creative}, {written printed instructional}, {written printed nonacademic}, and {written printed academic}, all with fairly small percentages of present perfects. The other is more heterogeneous, comprising two parts with considerably more present perfects, namely {written printed reportage} and {written printed persuasive}. This is interesting because, of all the written sub-registers, these are the ones one would intuitively regard as more closely related to speaking than the other written sub-registers (but see also below). The two sub-registers making up {spoken dialog} are extremely close to each other, as are the two sub-registers making up {spoken monolog}, and {spoken mix broadcast} is again an outlier with a markedly above-average percentage of present perfects.

To me, one of the most relevant advantages of this approach is that it allows the researcher to identify what appears to be relevant for present perfects in an exploratory bottom-up fashion. Rather than testing any particular preconception about which *level of granularity* (e.g., mode, register, sub-register) or *distinction at some level of granularity* may be relevant to present perfects, all variables included in one's corpus division can be included in comparisons to seek those which mark the largest differences, which would then be the natural starting points for more fine-grained follow-up analyses.

To summarise, I have shown how one can quantify the variability of one's own results by determining the range of *z*-scores for separate corpus parts and how this enhances the descriptive adequacy of one's results. Then, I illustrated how the variability can be explored on one chosen level of granularity, again using the *z*-scores. Finally, I demonstrated what a more sophisticated analysis on multiple nested levels of corpus organisation would look like.

Before the logic underlying this approach is explored a little further, it is worth emphasising here that this quantification of variability is not just interesting in its own terms or because it allows for assessing previous claims about the reliability of earlier studies. It also allows us to make a first step towards addressing another important issue, namely the issue of corpus homogeneity. In order to measure the homogeneity of a corpus with respect to some parameter, it is of course necessary to have divided the corpus into parts whose similarity to each other is assessed. While most previous studies of corpus homogeneity simply divided corpora into files and used word frequencies, the present approach allows for precisely quantifying the homogeneity of the corpus only on the basis of the parameter of interest, viz. present perfects. The most basic way to operationalise corpus homogeneity would be to use the range of values of the *z*-scores of the different corpus parts, but the following sections will introduce further, slightly more complex refinements of this approach.

3.2 More complex approaches to describing variability

I hope to have shown that the approach outlined in the previous section is a substantial improvement in terms of descriptive accuracy, ease of further exploration, as well as exploitability. However, this approach can still be elaborated, and the elaboration will be useful:

- to increase the level of precision with which the variability of the corpus is quantified;
- to address the threat posed by the data sparsity problem reflected by the sometimes small frequencies in the boxplots in Figure 1; and,
- to help the researcher decide on a principled basis how to split up the corpus at a given level of granularity in order to obtain

groups with the largest within-group similarity and the smallest between-groups similarity with respect to the present perfect's frequency.

As per the above, the refinements I shall outline are also very much inspired by exploratory data analysis using summary statistics and graphical methods. However, this section will also make use of the types of resampling approaches that are becoming more and more prominent what with today's computing power. The idea is to assess the variability coming with the results – and, thus, the homogeneity of the corpus – by recording the results one would have obtained if one had sampled just parts of the corpus which were investigated. Then, the more the results obtained from parts of the corpus resemble those from other parts of the corpus or those from the whole corpus, the less variable the results are and the more homogeneous the corpus is. Obviously, the more comprehensive the simulation, the more precise our results will be. The resampling method I will start with is permutation; Section 3.2.2 will be concerned with bootstrapping.

3.2.1 Exhaustive permutation

3.2.1.1 Using exhaustive permutation to assess the variability of a single summary statistic

Let me first show how resampling can be applied to summarising a single statistic such as the percentage of present perfects. In order to achieve the above goals with such a resampling approach, the same two steps that led to Table 3 were taken. As a third step, I then wrote an R script that generates all possible corpus parts one can form with s=13 sub-registers. That is, for each number s from 1 to 13, I formed all possible combinations of these s sub-registers. One obtains:

- 13 different corpus parts containing just one sub-register: {sub-register 1}, {sub-register 2}, {sub-register 3}, …, {s-r 11}, {s-r 12}, {s-r 13};
- 78 different corpus parts containing two sub-registers: {sub-register 1 and sub-register 2}, {sub-register 1 and sub-register 3}, …, {s-r 1 and s-r 13}, {s-r 2 and s-r 3}, {s-r 2 and s-r 4}, …, {s-r 11 and s-r 12}, {s-r 12 and s-r 13}; and,
- 286 different corpus parts containing three sub-registers: {sub-register 1, sub-register 2 and sub-register 3}, {sub-register 1, sub-register 2 and sub-register 4}, …, {s-r 1, s-r 2 and s-r 13}, {s-r 2, s-r 3 and s-r 4}, {s-r 2, s-r 3 and s-r 5}, …, {s-r 2, s-r 3 and s-r 13}, …, {s-r 10, s-r 11 and s-r 12}, {s-r 10, s-r 11 and s-r 13}, {s-r 11, s-r 12 and s-r 13}, *etc.* for all larger numbers of sub-registers.

This results in a total set of 2^s-1=8,191 differently-sized corpus parts. These are all the corpus parts one *could* have formed from the ICE-GB given the level of granularity of the sub-register.[8] The final step involves computing the relative frequency of present perfects in each of these 8,191 corpus parts, which results in a distribution of 8,191 percentages which can now be investigated with respect to homogeneity. Let us first start with the usual inspection of the frequency distribution, which is presented as a histogram in Figure 2.

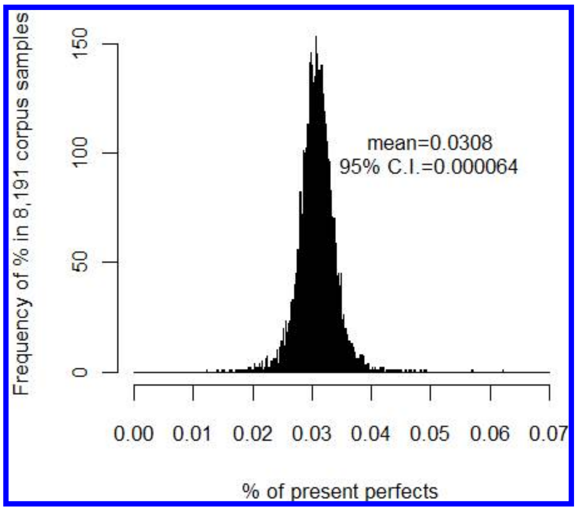


Figure 2: The distribution of frequencies of present perfects in all 8,191 sub-register defined parts of the ICE-GB

It is clear that the frequencies are distributed almost normally around the mean I established earlier. For the reasons mentioned above, the percentages cannot really be compared to Schlüter's figures and thus cannot address Schlüter's concern about reliability.[9] However, Figure 2 provides a clear indication of how much variability of present perfect frequencies the ICE-GB contains since it summarises all the frequencies of present perfects one could have found in the sub-registers of ICE-GB.

[8] Of course, any finer level of granularity is still possible – such as, for example, the individual corpus file – but the truly exhaustive enumeration of all corpus parts as used above would result in having to handle 2^{500}-1≈$3.273\cdot10^{150}$ corpus parts (this figure was computed using the GNU Multiple Precision Arithmetic Library at http://www.swox.com/gmp/) This exceeds contemporary computational limits, and will certainly do so for the foreseeable future. However, under Section 3.2.2 I will suggest a heuristic for incorporating finer distinctions.

[9] Of course it would be possible to compute the frequencies of present perfects per 1,000 words for the ICE-GB, but since these figures would suffer from the same kind of bias as those summarised by Schlüter, they would still not be comparable meaningfully.

Figure 2 also shows that while most of the sampled corpus parts yield very similar results, there is a variety of results that stray considerably from the mean. Thus, if such variation is found in a single corpus intended to be representative of British English in the 1990s, the variability found in many different corpora comes as no surprise. Finally, the corpus parts that deviate most strongly from the mean are of potential interest for further investigation.

3.2.1.2 Using exhaustive permutation to assess the sizes of pairwise differences

Again, this approach has more to offer than this. Above, we used boxplots to compare the different modes and (sub-)registers. Given the resampled corpus parts, we can now also perform some such comparisons on the basis of the 8,191 corpus parts. To illustrate, consider two distinctions often claimed to be very useful, those of spoken and written language samples, and printed and non-printed written data (see Biber, 1993: 245). On the basis of the data resulting from the permutation procedure, it is now possible to determine the importance of these or any other distinctions with the chosen granularity for the phenomenon under investigation, while representing the amount of variability at the same time. To this end, one could extract from the table of all 8,191 corpus parts all $2^5–1=31$ corpus parts containing only spoken language and all $2^8–1=255$ corpus parts containing only written language files and then compute all $31·255=7,955$ differences of these parts. The same can be done for the $2^2–1=3$ corpus parts containing only written non-printed language and all $2^6–1=63$ corpus parts containing only written printed language and the $3·63=189$ differences of these parts. Then, the pairwise differences of present perfect frequencies in these mutually-exclusive corpus parts can be compared to each other as in Figure 3.

Both distinctions play a noticeable role in accounting for a frequency difference of about 1 percent. More precisely, the proportion of present perfects in speaking differs from that in writing by approximately 1.1 percent, as does the proportion of present perfects in printed writing from that in non-printed writing. It is interesting to note that the more basic intuitive distinction between modes in fact results in a slightly smaller difference as that within one mode.

However, there is still room for improvement. First, the analysis of such differences is only easy to implement with pairwise comparisons, but one may often be interested in more distinctions at the same level of granularity and/or even more distinctions at other levels of granularity. Secondly, the finer the distinctions involved, the faster one runs out of parts to sample: there are many corpus parts containing only spoken or only written language to generate the left panel of Figure 3, but there are few individual parts to compare at the level of the sub-registers. Thirdly, if one simply tried to avoid that problem by adopting a finer-grained sampling, then the size of $2^{\text{number of parts}}$-1 may rule out this approach (cf. n. 8). Fourthly, the variation in each sub-register is left unaccounted for. The following section introduces an extension of the resampling approach that addresses these issues.

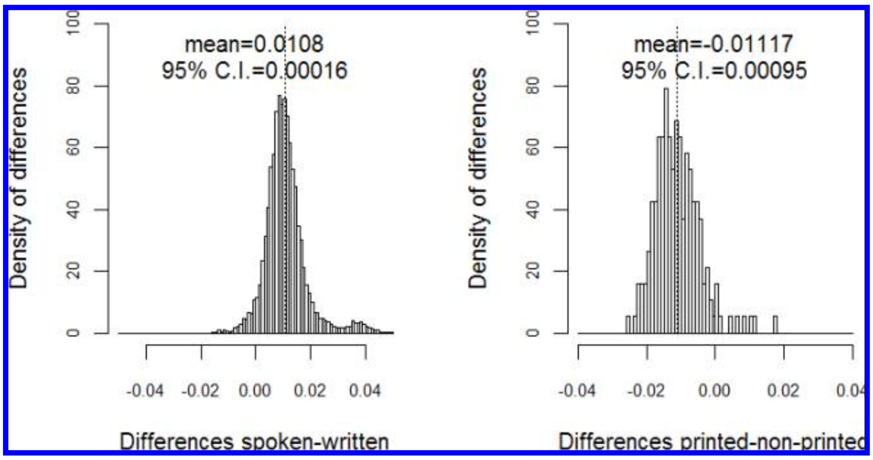


Figure 3: The distribution of frequencies of present perfects in speaking vs. writing and in printed vs. non-printed in the ICE-GB (Note: The vertical dotted lines indicate the overall means of the differences)

3.2.2 Bootstrapping

While the last section exhaustively permuted sub-registers, in this section I propose a bootstrapping approach on the basis of corpus files, i.e., at a level of granularity we have seen to be impossible to attain with the exhaustive permutation approach. Bootstrapping is the name of a statistical technique to estimate the distribution of a statistic by drawing random samples with replacement from the set of data points for which a particular statistic is desired.[10] From the list of all files and their percentages of present perfects (see Table 3), I drew 50,000 random samples with replacement from all files of each sub-register. That is to say, I drew 50,000 samples with replacement of size 100 from the percentages of all 100 files containing spoken private dialogues and computed and stored the mean for each of the

[10] Sampling elements out of a set with replacement, for example, differently-coloured balls out of an urn, means that whenever an item has been drawn out of the urn, the result of the draw is noted and the item is placed back into the urn before the next item is drawn from the urn.

50,000 samples. Then I drew 50,000 samples with replacement of size eighty from the percentages of all eighty files containing spoken public dialogues and computed and stored the mean for each of the 50,000 samples, *etc*. I thus obtained a vector of 50,000 means of present perfect means for each sub-register.

One minor problem associated with the boxplots was that while the method could test the importance of all distinctions at different levels, researchers must still try to work out the optimal grouping at each level for themselves. More specifically, the third panel of Figure 1 requires the researcher to decide whether to group {letters} together with {scripted} or with {persuasive}. Since this decision involves somehow weighing medians and dispersion, different analysts may reach different decisions. However, as each sub-register is now characterised by a vector of 50,000 means, one option is to derive the groupings directly from the data by means of, for example, a hierarchical agglomerative cluster analysis (see Kaufman and Rousseeuw, 1990). Figure 4 is the result of such an analysis (with Euclidean distances as the similarity measure and average distance as the amalgamation rule).

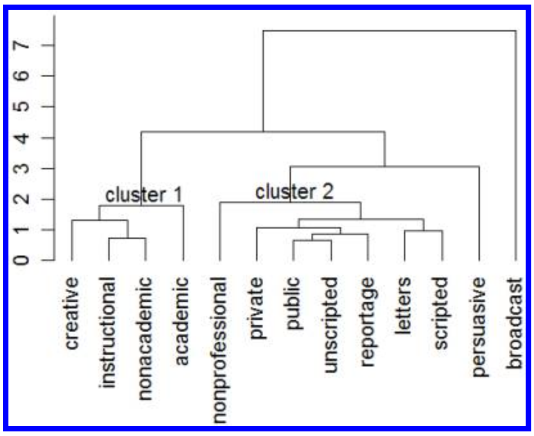


Figure 4: Dendrogram of a hierarchical agglomerative cluster analysis on 50,000 bootstrapped means of all sub-registers of the ICE-GB

The cluster analysis provides us with several subparts which a researcher should recognise if he or she is interested in exploring the variability of the present perfect in the ICE-GB. Cluster 1 contains printed, elaborated and mostly information-transmitting writing with a smaller-than-average proportion of present perfects. Cluster 2, by contrast, contains sub-registers which are characteristically more interactive and less formal and which exhibit a fairly average degree of present perfects. Note that nearly all spoken sub-registers are found in this cluster. {Spoken printed persuasive} is somewhat exceptional: it is still connected to cluster 2, but at some distance. This ties in very well since persuasive writing – in this case many political press editorials – approximate direct conversation in their attempt to get the reader to adopt a particular opinion, which would be made more difficult when attempted with a formal and stilted writing style. Finally, the outlier of {spoken mix broadcast} is amalgamated into the tree at the very end of the amalgamation process, again testifying to its peculiarity.

These findings are interesting in two respects. Methodologically, they provide the researcher interested in English present perfects with a data-driven and objective categorisation of the groups of data in his or her corpus. The categorisation is objective because all the steps to arrive at these groups, with the exception of the choices of the similarity measure and the amalgamation rule, do not involve any (potentially biased) human decisions. As such, the results of the cluster analysis also again bear on the issue of corpus homogeneity: the corpus parts from which the bootstrapping was conducted were defined using the corpus's sub-registers and the dendrogram reflects the corpus parts most homogeneous internally and most heterogeneous externally on the basis of the parameter of interest.

Note that, exactly because this bootstrapping approach is not limited to any particular level of granularity, it is the solution to the memory overflow problem that would arise with the exhaustive permutation approach. For example, one could also have done the same at a finer level of granularity. One could have taken each file with its *n* verbs, drawn 50,000 samples with replacement from each file, computed the number of present perfects out of the *n* verbs obtained in the 50,000 samples and then stored these 50,000 values for each file in one vector for each file. This list of vectors would then constitute the input to the cluster analysis to determine which (groups of) files are most similar to each other *etc*. Thus, the bootstrapping approach allows for identifying homogeneous groups at whatever level of granularity one is interested in without imposing any distributional assumptions (registers, sub-registers, files, parts of files, speakers, *etc*.).

Conceptually, these findings are interesting because, although only just a single and very simple variable was involved, the clusters conform well to two of Biber's most important, linguistically-defined factor dimensions, 'informative vs. interactive' and 'elaborated vs. situation-dependent'. I am not at all suggesting that we can do away with the so much more comprehensive kind of analysis used by Biber and his colleagues: the reverse is true. However, it shows that the method proposed here is not just objective but also a useful heuristic for identifying sources of variability that tie in with much more elaborate studies. This, I interpret as considerable support for the method (see Schlüter, 2002: 111 for a similar claim concerning how present perfect frequencies can distinguish text types).

This case study has been restricted to the methodologically simplest summary statistic: percentages. To demonstrate that the methodology proposed is applicable to any kind of statistic derived from corpus data, the next section will be concerned with a shorter, but statistically more complex example.

4. Case study 2: predictability and persistence of a constituent order alternation

In this section, I will be concerned with the influence of structural persistence on, and the predictability of, constituent order alternations and the way in which resampling allows for identifying corpus parts requiring more refined analysis.

Structural persistence refers to the fact that speakers/writers are more likely to use a syntactic pattern if they have just used a similar syntactic pattern before. The classic study of this phenomenon is probably Bock (1986). Under the guise of a memory task, subjects first repeated prime sentences having one of two alternating structures: the transitivity alternation (i.e., active vs. passive sentence form) or the so-called dative alternation (i.e., the ditransitive, or double-object, construction vs. the prepositional dative with *to* and *for*). Then, the subjects described semantically-unrelated pictures allowing both syntactic alternatives. Bock found that subjects preferred to formulate a description whose syntactic structure corresponded to that of the prime sentence. While the mechanism underlying this phenomenon is still hotly debated, persistence has been replicated in many studies and languages.

In this section, I will discuss a second short case study concerned with structural persistence from a corpus-based perspective. The main point of this section is to show that the kind of approach outlined in Section 3 does not only apply to mere percentages, but also extends naturally to analyses that are statistically more complex. The example I will use here to make this point is that of particle placement in English, i.e., the constituent order alternation of transitive phrasal verbs exemplified in (1).

1. a) John picked up the book.
   b) John picked the book up.

This phenomenon has been investigated in many studies; see Gries (2003) for a comprehensive overview. Some recent studies have also investigated this alternation with respect to the phenomenon of structural persistence. Gries (2003: Section 6.4) mentions persistence only very briefly. He investigated whether the instances of the verb-particle construction in his corpus were primed by another instance of the verb-particle construction in the preceding three clauses and found a significant correspondence of verb-particle constructions in the two nearby clauses.

Two studies looking at persistence in much more detail are Gries (2005a) and Szmrecsanyi (2005). Gries conducts multifactorial analyses of persistence of the dative alternation and particle placement in the ICE-GB, using 3,003 instances of the dative alternation and 1,797 instances of the verb-particle construction. Gries finds strong support for persistence in both alternations, with some effects and effect sizes closely resembling those obtained in experimental studies. For example, he finds that priming effects are stronger in speaking than in writing, if the verb in the prime and the target is identical, and that the distance between prime and target is inversely, nonlinearly related to the strength of the persistence effect. Finally, he notes a strong verb-specific bias: not all verbs exhibit persistence effects to the same degree – some verbs prefer to stick to one construction irrespective of whether they are 'primed' or not, while others readily undergo 'priming effects'.

In a similar study, Szmrecsanyi (2005) investigates particle placement in the Freiburg English Dialect Corpus (FRED). He investigates how several 'traditional variables' such as definiteness, length, news value, literalness of the direct object *etc*., influence the choice of constituent order in 1,048 verb-particle constructions. In addition, he also includes in his analysis the impact of the dialect area, the previous syntactic pattern in a verb-particle construction and its distance to the target construction, and sentence length (as a proxy of sentences' complexity). His results provide strong support for the relevance of most traditional factors as discussed in Gries (2003) and also corroborates Gries's (2005a) results that persistence is positively correlated with identical lemmas and short prime-target distances. Szmrecsanyi also discusses a variety of interesting interactions of factors with persistence.

While Gries (2005a) and Szmrecsanyi (2005) go beyond much previous corpus-based work on persistence, they do not investigate to what degree the findings they present are robust in the sense that they would be obtained through different corpora or different corpus parts. In this study, I will briefly exemplify this issue on the basis of data investigated in Gries (2005a), verb-particle constructions from the ICE-GB. However, I will restrict myself to (i) a much smaller number of factors and (ii) prediction accuracies from binary logistic regressions. The objective is to identify the corpus parts which yield the worst prediction accuracies and are thus the most rewarding starting points for further analyses or refinements.

As a first step, I extracted all examples of the verb-particle construction from the ICE-GB. These were entered into a table (in their order of occurrence) and each instance was annotated with respect to the following set of variables:

- the sub-register in which each instance was found, using the classes from above;
- the verb and the particle making up the transitive phrasal verb instantiating one of the two constructions;

- whether the verb lemma or the particle was the same in a consecutive prime-target pair;
- the distance between two consecutive verb-particle constructions in terms of parse units of the ICE-GB; in the analysis, the natural log of this distance was used;
- a verb attraction measure quantifying which syntactic pattern the transitive phrasal verb prefers to occur in and how strongly it does so: high negative values represent a statistically significant association to the V-Part-DO order while high positive values represent a statistically-significant association to the V-DO-Part order;[11] and,
- the persistence factor: if the construction was neither the first nor the last in one corpus file, the syntactic pattern that was chosen in the previous verb-particle construction and that of the subsequent verb-particle construction. When a construction was the first or last of the file, no data were coded for a previous or subsequent occurrence respectively.

As a second step, I first wrote an R script that generated all 8,191 possible groupings of sub-registers and computed 8,191 binary logistic regressions for all verb-particle constructions in each grouping of sub-registers. A binary logistic regression is a regression technique with which one attempts to predict the outcome on a dependent variable with two levels on the basis of a set of categorically and numerically independent variables. In these logistic regressions, the choice of construction – V-Part-DO vs. V-Part-DO – was the dependent variable, while the other variables from above constituted the independent variables; for simplicity's sake, interactions were not included in the model. More simply, what is at issue in the present case is trying to predict which of the two verb-particle constructions a native speaker will choose, given that one knows the verb, the particle, the last-used verb-particle construction *etc*. The output for such binary logistic regressions usually includes:

- statistics that serve to evaluate the overall utility of the model: Nagelkerke's R^2 (i.e., a correlation coefficient), the prediction accuracy achieved on the basis of the model, a likelihood ratio chi-square, *etc*.; and,
- odds ratios to assess each independent variable's contribution to the prediction.

The statistics outputted by the R script include all these usual results but I will only focus on Nagelkerke's R^2, the prediction accuracy achieved on the

[11] This association was computed using distinctive collexeme analysis, a method from the family of methods referred to as collostruction analysis; cf. Section 5, Gries and Stefanowitsch (2004), and Gries (2005a) for details.

basis of the model, and most importantly the difference between the *observed* prediction accuracy and that *expected* by chance, which will be referred to as prediction improvement. The results for these analyses are summarised in Figure 5.

As Figure 5 shows, the present data resulted in slightly lower R^2's and slightly lower prediction accuracies than those reported in Szmrecsanyi (2005). However, since that study looked at many more variables than could be dealt with here, this was to be expected. Also, the persistence factor in Szmrecsanyi (2005) has a stronger impact on the predictability than it does here. This is probably due to the fact that Szmrecsanyi only investigated spoken data, where persistence is stronger than in the written data that I included. It is also interesting to notice that the effect size of the verb attraction measure (not represented graphically) usually outperforms that of the persistence factor, yet the opposite was true in Szmrecsanyi's study. However, while the overall summary results are relatively similar, it is also obvious that not all corpus parts allow for the prediction of constructional choices equally well, as is shown by the range of the whiskers and the outliers in Figure 5.

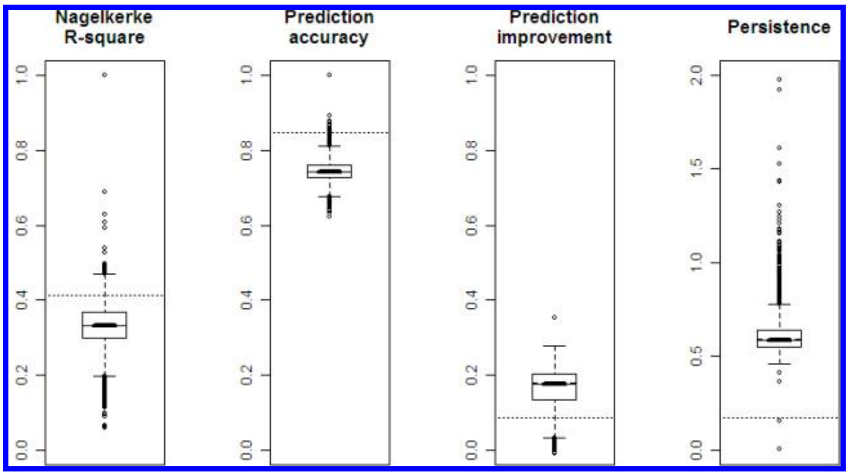


Figure 5: Boxplots of Nagelkerke's R^2 in this study (left), obtained prediction accuracies (left centre), the improvement of the prediction over change (right centre) and odds ratios for the factor of persistence (right) for particle placement in all 8,191 sub-register partitions of the ICE-GB
(Note: Dotted lines represent Szmrecsanyi's (2005) findings)

To explore the merits of the resampling approach and determine where the variability in the results comes from, I inspected the prediction improvements for (i) the thirteen sub-registers and (ii) the thirteen most frequent particles in the transitive phrasal verbs (those with $n \geq 20$: *away*, *back*, *down*, *forward*, *in*, *off*, *on*, *out*, *over*, *round*, *through*, *together*, *up*).[12] The upper panel of Figure 6 plots the prediction improvements (on the *y*-axis) obtained in different corpus parts against the number of verbs (on the *x*-axis) which entered the model; this latter complication is necessary to show that not all bad model fits are just due to small corpus sizes.[13]

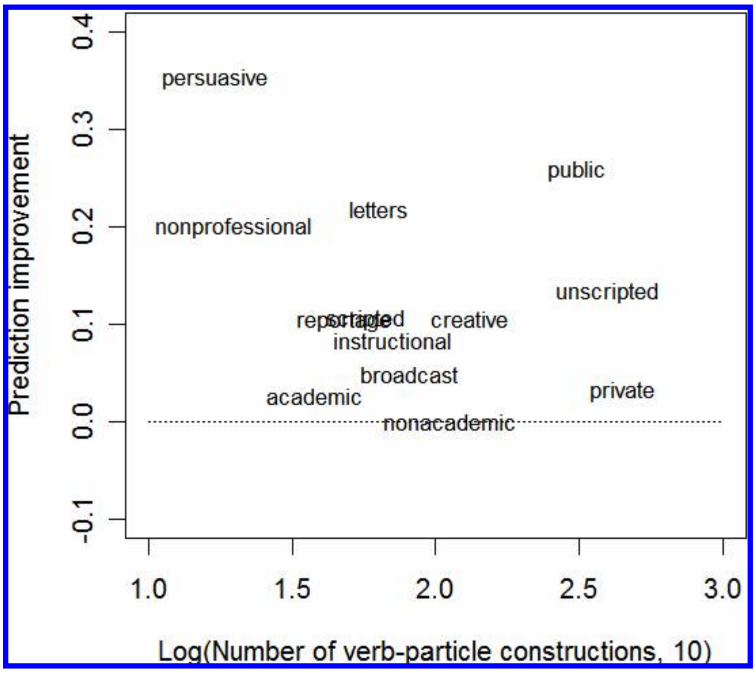
The most striking result is that sub-registers differ enormously in terms of how well native speakers' constructional choices can be predicted: in several kinds of elaborated, informative writing (e.g., non-academic writing), constructional choices can hardly be predicted better than by chance. These parts point to areas requiring further investigation. On the other hand, three of the four sub-registers with the highest prediction improvements (up to approximately 35 percent) are from less formal, more interactive writing: {written printed persuasive}, {written nonprinted nonprofessional} and {written nonprinted letters}. This is especially interesting because most of these samples are rather small in size.

The lower panel of Figure 6 presents the same results for the corpus parts defined in terms of the particles. These are also noteworthy because they first show that particles differ strongly in terms of how well their position with respect to a verb can be predicted. Contrary to what we may assume, however, the prediction improvements again do not simply increase proportionally to the sample sizes. In fact, the two particles that are the most frequent by far, yield just intermediate prediction improvements. Interestingly, the constituent orders of phrasal verbs with the two largely directionally-used particles *round* and *back* cannot even be predicted beyond chance accuracy.

In sum, this section has demonstrated, if only briefly, that the present approach can also be applied to quantitative results that are more complex than mere percentages, namely to complex multifactorial coefficients. Also, it was shown how splitting and resampling identifies corpus parts that differ in terms of predictability of a particular phenomenon, that fall into groups that correlate with Biber's results from much more comprehensive studies, and that provide the most natural starting point for subsequent analysis.

[12] In order to shorten an already long paper, I will not discuss results for the differences between the 8,191 corpus parts of the sub-registers and those of the particles.

[13] Strictly speaking, it would be more useful not to use the prediction improvement deriving from the model as such since then the prediction accuracies are based on the same data from which the model is derived. Thus, ideally one would identify the difficult-to-predict corpus parts using cross-validation. While, unfortunately, neither Gries (2005a) nor Szmrecsanyi (2005) report cross-validated prediction accuracies, the present data allow for generating these quite easily. For example, on the level of sub-registers, one could just take each of the 8,191 corpus parts, derive a logistic regression model from it, use the models to predict each of the other corpus parts, and record the degree to which the model allows for correct predictions.



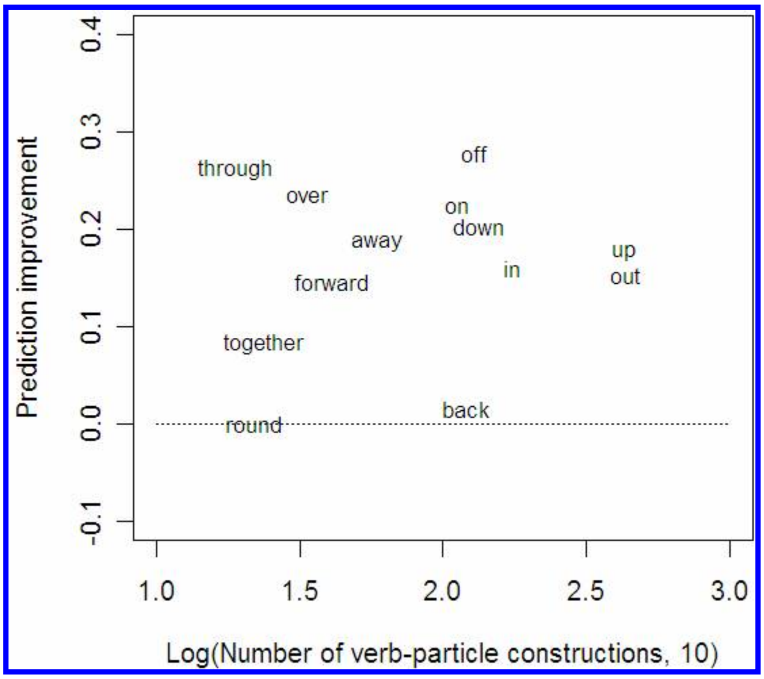


Figure 6: Prediction improvements of particle placement in mutually-exclusive corpus parts of the ICE-GB
(Note: Dotted lines represent chance prediction accuracy)

5. A more general approach to corpus homogeneity: ditransitives in English

We have already seen in Section 3 that the resampling approach allows us to quantify the homogeneity of a corpus with respect to any parameter of interest in various ways: *z*-scores, histograms, boxplots, dendrograms. However, the width of a histogram or the range of *z*-scores of sub-registers can serve as a first simple heuristic only. And while the cluster analysis can determine the most interesting groupings of one's corpus parts, it is difficult to imagine how one would use it to compare different corpora or different divisions of the same corpus (see below). Thus, a more rigorous and comparable procedure would be desirable. In this section, I will outline an idea for such an approach, using as an example the method of collexeme analysis.

Collexeme analysis is one recent corpus-based approach from the family of methods called collostructional analysis and is used to investigate the semantics of (argument structure) constructions by identifying the words that are (significantly) associated with syntactically-defined slots of constructions. It is, thus, similar to collocational studies and assumes a Construction Grammar approach to language (see, for example, Goldberg, 1995), according to which lexis and grammar form a continuum of elements and where the degree to which elements tend to co-occur is governed by their degree of semantic similarity and compatibility (see Stefanowitsch and Gries (2003) for details). To investigate the semantics of an argument structure construction such as the English ditransitive, the following steps must be taken:

(i) look for all examples of the ditransitive (which often involves semi-manual coding);

(ii) retrieve all words occurring in a particular syntactic slot of the ditransitive (usually the main verbs) as well as their overall frequencies in the corpus to generate for each verb a 2×2 co-occurrence table of the kind represented in Table 5; and,

(iii) from each such 2×2 co-occurrence table, one computes a measure of association strength (called collostruction strength) to determine (a) the *direction* of the co-occurrence, i.e., whether the construction and the word co-occurs more or less frequently than expected, and (b) the *strength* of the more-or-less-frequent-than-expected co-occurrence: the higher the value, the stronger the effect.

As previous work has shown, the semantics of constructions can usually be read off from the words most strongly attracted to the slot in question in a given construction; these words are referred to as *collexemes*. For example, the top-ranked verb collexemes obtained for a collexeme analysis of the ditransitive by Stefanowitsch and Gries (2003: Section

3.2.2) are summarised in Table 6, and this clearly shows that both the sense of transfer, which is usually associated with this construction, and the many semantic extensions the ditransitive is argued to have, are strongly reflected in the top collexemes.

	Construction *c*	¬ Construction *c*	Row totals
Word *w*	*a*	*b*	*a*+*b*
¬ Word *w*	*c*	*d*	*c*+*d*
Column totals	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*=*N*

Table 5: Schematic 2×2 co-occurrence table for the statistical analysis of collexemes

Verb	*Collostruction strength*	*Verb*	*Collostruction strength*
give	infinity[14]	*allow*	9.95
tell	126.8	*lend*	8.55
send	67.14	*deny*	8.35
offer	48.48	*owe*	7.57
show	32.65	*promise*	7.49
cost	21.95	*earn*	6.67
teach	15.36	*grant*	5.88
award	10.87		

Table 6: Top fifteen collexemes of the ditransitive construction in the ICE-GB (data from Stefanowitsch and Gries, 2003)

While collexeme analysis has been applied to a sizeable number of constructions and phenomena, this previous work, in common with most other corpus-linguistic studies, has largely focused on the overall results obtained for a particular corpus. Thus, what has been neglected is the degree of variability within the corpus, i.e., the homogeneity of the corpus investigated with respect to the construction in question. In this section, I will show how to assess the homogeneity of the corpus with respect to the parameter of interest – the attraction of verbs to the verb slot of the ditransitive construction – using a so-called homogeneity plot. As mentioned before, quantifying corpus homogeneity requires splitting up the corpus into parts to be compared, and in this section I will exemplify the

[14] This value was larger than R's computing limit on the computer used.

concept of homogeneity plots by splitting up the corpus along the by now familiar parameters of files, sub-registers and registers.

First, I wrote R scripts that (i) retrieved all ditransitives from the ICE-GB, (ii) retrieved all verbs and their lemma frequencies in ditransitives and elsewhere, and (iii) stored the files in which each verb occurred in a ditransitive. It turned out that in all 500 files of the ICE-GB eighty-eight ditransitive verbs occur. Secondly, I computed the collostruction strength of each verb to the ditransitive for each of the 500 files, ultimately obtaining a 88×500 table in which each cell contained the collostruction strength of the verb of the respective row in the file of the respective column.[15] At the same time, these steps were also performed with the thirteen sub-registers and the five registers of the ICE-GB, yielding two analogous tables with dimensions of 13×500 and 5×500.

The final step consisted of evaluating these two-dimensional tables with principal components analyses. The logic underlying this approach is as follows. A principal components analysis is a data reduction method that takes as input a table with, in the case of files, 500 columns. It then tries to represent as much of the variance, *v*, contained in this table as possible but reduce the number of columns. This reduction is achieved by detecting intercorrelations of columns of the table and summarising sets of highly-correlated columns as so-called principal components or factors. Thus, the more structure there is in the table, the fewer principal components/factors will be needed to summarise as much of the information contained in the original table as possible. Theoretically, the smallest number of principal components/factors would be one, meaning that only one principal component/factor would be needed to represent all the variance within the table because all columns are so highly correlated. The largest number would be the original number of columns, in which case all the variance was represented, but all columns turned out to be so different from each other that no reduction was possible.

From this it follows that we can represent the homogeneity of the corpus with respect to the feature under consideration (as represented by the complete table) by plotting the amount of variance, *v*, in the table that is represented against the number of principal components/factors needed to represent it.[16] The more variance one can explain with few principal components/factors, the more structure the table contains. Thus, the larger the slope of the line, the more homogeneous the corpus is with respect to the chosen division of the corpus (i.e., files, sub-register, register, *etc.*). By

[15] Most previous collexeme analyses used -log10 ($p_{\text{Fisher-Yates exact test}}$) as the measure of association strength. However, *p*-values are dependent on sample sizes so the files, sub-registers, and registers of the ICE-GB, which differ in their size, would distort the results. Thus, for this study, I used $\log_{10}$ odds ratio as a measure of association strength that is not dependent on the sample size the way *p*-values or χ^2-values are.

[16] This representation is quasi a reversed scree plot. In order to be able to compare analyses with different numbers of columns, however, it is useful not to use the absolute number of principal components, but the percentage of principal components out of all columns.

contrast, if the slope of the line approximated unity, each separate factor extracted only explained about as much variance as one of the original columns and the corpus would be very heterogeneous.

Before we look at the results for the three different corpus divisions, let us briefly formulate what one would expect to find if the method worked. To me, reasonable expectations would be the following:

- The homogeneity of the corpus divided according to files should be relatively high because there is little reason to assume that a division of the corpus according to something as arbitrary as files would introduce much systematic variability. This would rule out explaining a lot of variance with few factors.
- The homogeneity of the corpus divided according to sub-registers should be relatively low because sub-registers as defined here are clearly-defined situational genre types. This should be reflected by substantial differences between columns and, thus, a low ratio of explained variance to number of factors.
- The homogeneity of the corpus divided according to registers should be fairly low and probably even lower than that of the sub-registers because the registers will lump together what is different on the level of sub-registers, introducing further heterogeneity.

Consider now Figure 7 for the corpus homogeneity plots resulting from the three analyses. Panel 1 shows the results based on the file-based division of the corpus.[17] The horizontal and vertical dotted lines are intended to facilitate the recognition of how much variance is explained with how many factors (an example will be discussed below). In an analogous fashion, Panel 2 and Panel 3 show the results for the sub-registers and registers respectively. Thus, the panels give a first impression of how corpus homogeneity can be assessed depending on which division of the corpus is chosen for the analysis.

All the expectations are borne out by the results. The extraordinary steepness of the slope in the homogeneity plot in Panel 1 indicates that the ICE-GB is quite homogeneous in terms of how particular verbs are attracted to the ditransitive construction across the files. For example, 435 files entered into the analysis and contributed variance to the overall table. However, forty-four factors (=10 percent on the *x*-axis) already explain more than 90 percent of the variance in the table (≈90 percent on the *y*-axis) *etc*. The other two homogeneity plots show that, as expected, both sub-registers and registers are much less homogeneous with respect to the same

[17] Since sixty-five files contained no ditransitives, these were omitted from consideration. If one wanted to include them, arguing that this lack of ditransitives ought to be represented as well, the upper bound of the plot in the left panel of Figure 7 would be at 0.87 (=435/500), but the slope of the line would be nearly identical to the one represented here so, for reasons of space, I omitted this homogeneity plot.

phenomenon. While this is not easily discernible from the graphs, inspecting the figures entering into the plots show that the amount of explained variance of the sub-registers is always as large as, or slightly larger than, that of the registers, conforming to the expectation perfectly.

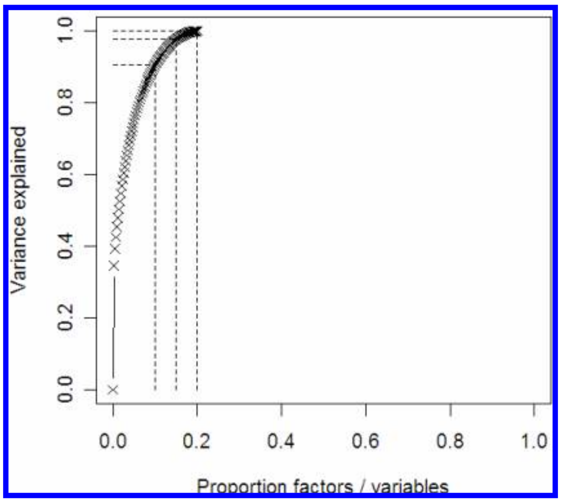


Figure 7 (Panel 1: *File-based results*): Corpus homogeneity plots: explained variance vs. percentage of principal components/factors

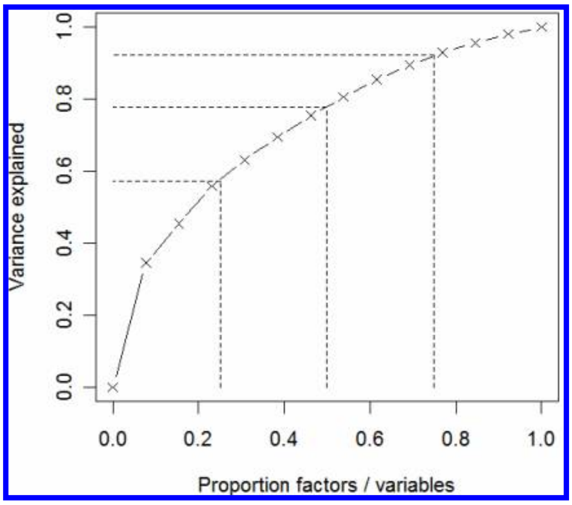


Figure 7 (Panel 2: *Sub-register-based results*): Corpus homogeneity plots: explained variance vs. percentage of principal components/factors

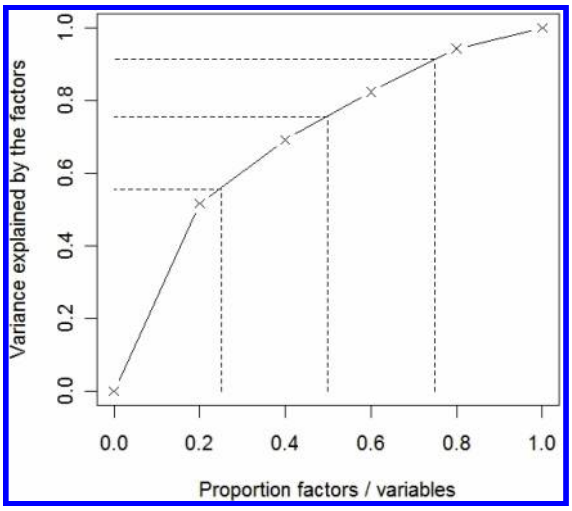


Figure 7 (Panel 3: *Register-based results*): Corpus homogeneity plots: explained variance vs. percentage of principal components/factors

In order to support the proposed method further, however, I tested it in another way. It was shown that the slope of the line representing the ratio of the percentage of factors vs. the amount of variance explained by the factors increases when the corpus is more homogeneous with respect to the chosen divisions. It was also shown that if the corpus is divided into thirteen meaningful situationally-defined registers, homogeneity is decreased, since the parts reflect dimensions introducing meaningful detectable variance. If one now generated thirteen parts *randomly* – that is, in a way that does not introduce *meaningful* detectable variance – then the homogeneity of the corpus should increase again. This is because, if the separation of the corpus is truly random, it should not give rise to structure that a principal components analysis could detect. To test this hypothesis, I wrote an R script to:

(i) divide the corpus into thirteen random parts;
(ii) compute a complete collostruction analysis of the ditransitive for each of the thirteen random parts and store all results in a table with thirteen columns (the sub-registers) and eighty-eight rows (the eighty-eight verbs); and,
(iii) do a principal components analysis on this table to see how much variance is explained by how many factors.

The results of such a procedure are presented in Table 7. The first column gives the number of factors identified by the principal components analysis. The second column translates the number of factors into a percentage out of thirteen original columns (see n. 16 above); thus, these are just the values of 1 divided by 13. The third column provides the amount of explained variance when the corpus was divided into meaningful sub-registers (i.e., the values on the *y*-axis of Panel 2 in Figure 7) and the final column gives the amount of explained variance when the corpus was divided randomly.

Number of factors	*% of original columns*	*Expl. var.* (*sub-registers*)	*Expl. var.* (*random*)
1	0.0769	0.3379	0.4274
2	0.1538	0.4485	0.5320
3	0.2308	0.5534	0.6091
4	0.3077	0.6269	0.6811
5	0.3846	0.6916	0.7436
6	0.4615	0.7526	0.7944
7	0.5385	0.8034	0.8362
8	0.6154	0.8524	0.8744
9	0.6923	0.8936	0.9084
10	0.7692	0.9285	0.9375
11	0.8462	0.9552	0.9608
12	0.9231	0.9806	0.9822
13	1.0000	1.0000	1.0000

Table 7: Amounts of explained variance: a meaningful vs. a random corpus division

With the exception of the last row, (which, by definition, is 1 in both cases), the amount of explained variance in the randomly divided corpus is higher than the amount of explained variance in the sub-register-based parts in all rows. The sums of all differences – Σ(column 4 - column 3) – add up to 0.4625. Since this is what was predicted, this finding provides further support for the method. However, lest someone raise further objections to the random sampling results, my R script did the above three-step simulation process 100 times. As a result, I obtained 100 vectors. Each of these corresponds to the final column of Table 7, but contains different values since the values are dependent on the random division of the corpus into 13 parts. I then computed the differences of the thirteen values of these 100 vectors from their corresponding value in the third column of Table 7 (i.e., 0.4274 – 0.3379, 0.532 – 0.4485, *etc.*) to see how often the amount of explained variance in the randomly-divided corpus was smaller than that of the sub-register-defined corpus, which would undermine the proposed method. The results are unanimous: all overall sums of differences are positive, many are quite high, as is shown in Figure 8, and none of the 1,300 comparisons were negative.

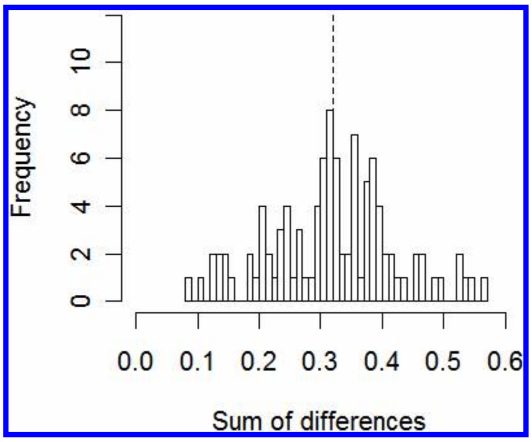


Figure 8: Histogram of the sums of vector differences (Note: The dotted line represents the overall mean)

With 100 sums of differences (of which the aforementioned value of 0.4625 is just one), one can even summarise the homogeneity of a corpus in a single index, namely the mean of these sums (or 0.3198 in this case). Looking at this mean of the many random divisions also has the advantage that this index is not easily inflated by outliers because any particularly peculiar random sampling can influence the overall mean only marginally and will be easily detected by a histogram such as that in Figure 8. All of this would amount to defining corpus homogeneity as in (2).

2. The homogeneity of a corpus with respect to a particular phenomenon and a particular level of granularity is proportional to the average amount of structure detected by a principal component analysis that exceeds that of a corpus that was divided into the same number of parts randomly.

This index is just one suggestion of how to explore this method further. Another would be to take the size of the area between the plotted line and the main diagonal as an index –. I interpret the results as strong support for the proposed method although further refinements of the method, as well as across corpora testing, may well be necessary.[18] For

[18] A refinement of this way of analysis may be necessary to account for the fact that, if there are more columns in a table, then intercorrelations boosting the amount of explained variance may arise by chance. However, this technicality does not undermine the general logic underlying the approach.

example, while this section has used principal components analysis, alternative techniques may also be conceivable. One possibility would again be to use cluster-analytic techniques. For example, agglomerative nesting as implemented in R yields an agglomerative coefficient which indicates the amount of structure the cluster analysis found in the data: the agglomerative coefficient ranges from 0 to 1, and the higher it is, the more structure there is in the table. Thus, as above, the more structure there is in the table, the less homogenous is the corpus. However, since the agglomerative coefficient increases with the number of elements that are clustered, it is not ideally suited to comparing data from differently divided corpora unless this tendency could be corrected.

Another possibility would be to use an information-theoretic approach. For example, Benedetto, Caglioto and Loreto (2002) used data compression to show how clustering compression ratios allowed for constructing surprisingly accurate language-family trees. In our case, it may be possible to measure the amount of structure in the table using compression ratios: the more the table can be compressed (relative to a reference table that figures the size of the table into the equation), the more heterogeneous is the corpus. In combination with this, or as an alternative, an approach based on entropy is also conceivable (see Sekine, 1997).

A further way of exploiting the present approach would be to look at the results of the principal components analysis for the non-random distinctions in more detail. For example, Gries (forthcoming: Section 4) applies a principal components analysis to verbs' collostruction strengths to the ditransitive and shows how the factorial structure (in terms of *Eigenvalues*>1) reflects four clearly distinct corpus parts. The interesting aspect of that result is that the corpus parts thus obtained cut across all levels of corpus categorisation – something which a human analyst would probably rather not do, but which the analysis nevertheless shows to be the most useful division of the corpus. For the moment, however, I must leave the exploration of these and other possibilities for future work.

6. Summary and conclusions

Beginning with data used by Schlüter to address the issue of reliability of corpus findings, I noted that discussing the reliability of any particular statistic is difficult without a precise indication of both the internal and external variability of the data sets in question. From the fact that corpus homogeneity/variability involves making decisions concerning the parameter of interest as well as the desired level of granularity, I illustrated several ways to address these issues. Specifically, I proposed ways to determine, (i) how large the differences between results of different corpus parts are on each level, and (ii) which level introduces the largest differences between results. I proposed that the results of both of these issues constitute important quantitative information in their own right as well as objectively identify from the data ideal starting points for subsequent analysis in a bottom-up fashion. That is, the proposed methods involve splitting up one's corpus on the basis of parameters of different degrees of granularity as well as splitting it up into parts and then using permutational and/or bootstrapping approaches to investigate central tendencies, frequency distributions of differences of mutually-exclusive corpus parts and homogeneous groups in one's data. Finally, I proposed to quantify corpus homogeneity using multivariate exploratory data analysis techniques.[19]

[19] I would like to take up one point raised by one anonymous reviewer, who criticised this paper because I did not use or at least mention *t*-tests and ANOVAs, which address the problem of within- and between-groups variability. Several comments are in order. First, it is well known that using an *F*-test to test the significance of differences between means presupposes that, (i) the error components (the deviation of an individual data point from the sample mean) are distributed approximately normally, (ii) the variances of the error components must be identical in all populations from which samples are investigated, and (iii) each error component is independent of the other error components. When these assumptions are met, the use of parametric statistics is unproblematic. However, as Pedersen, Kayaalp, and Bruce (1996: 458; emphasis added) state:

> If the statistic used to evaluate a model has a known distribution when the model is correct, that distribution can be used to assign statistical significance. […] As discussed so far, this distribution can be approximated when certain assumptions hold. *The problem is that these assumptions are frequently violated by the data found in NLP.*

This is also true of our present case: there is no reason whatsoever to assume that the absolute or relative frequencies of the linguistic elements under consideration meet these conditions, and there is probably even less reason to assume that all the statistical parameters I investigate above (Nagelkerke's R^2, log odds ratios, *etc.*) do so. All this is so widely recognised that it is found in the current standard textbook references (e.g., Manning and Schütze, 2000: 169, 172) and responsible for the fact that *t*-tests and ANOVAs – in fact, most parametric techniques – have not been prominent in the relevant literature at all. Note especially that the recommendation given by Pedersen, Kayaalp, and Bruce, 1996: 458) is exactly the one adopted in this paper:

> An alternative to using an approximation to the distribution of goodness of fit statistic is to define its exact distribution by enumerating all elements of that distribution or by sampling from that distribution using a Monte Carlo sampling scheme.

Since the resampling approaches used to determine variability and homogeneity in the present paper make no such assumptions, they are much more robust than the suggested parametric alternatives. Secondly, while being more robust in the above sense of distributional assumptions, the kind of resampling approaches used in the present study are also more robust in the sense that, for example, individual outliers can be identified straightforwardly and have much less of a chance to bias results.

Thirdly, some of the techniques used here also facilitate the analysis of the results on multiple levels of granularity much more intuitively. Boxplots *etc.* would not allow for objectively determining the kind of hierarchical groupings that are so easily recognisable from a cluster-analytic dendrogram because researchers may differ in their interpretation of the visual displays. Alternatively, one could do much more complex and less intuitively understandable *post hoc* analyses of means (using, for example, Newman and Keul's or Duncan's range tests), but while these tests would help identifying groups in the data, they would not tell how and at what distances these groups are related. Finally, a trival point: the fact that *t*-tests and ANOVAs are the parametric techniques to investigate between-groups and within-groups variance does not prove that other statistical methods cannot also yield interesting results. It is for these reasons that, in spite of their appeal at a superficial glance, *t*-tests and ANOVAs do not enjoy a central status here.

I submit that this kind of approach – irrespective of its exact method of implementation – has a variety of advantages over much previous corpus-based work in general as well as work on corpus homogeneity. First, it provides an indication of the dispersion of the statistical parameter that is reported. This result is more informative and robust than any single confidence interval, standard error, *etc*. and so enormously enhances the descriptive adequacy of the results, as well as the potential to generalise.

Secondly, it allows one to apply flexibly the methods to parameters of all levels of statistical sophistication: simple frequencies, percentages or conditional probabilities, association measures, regression weights, *etc*.

Thirdly, it allows for objectively identifying (i) corpus parts exhibiting noteworthy results on one's parameter of interest (e.g., outliers within, or tails of, distributions of differences); and (ii) the major determinants of variability within whatever set of corpus divisions appear sensible (e.g., register or sub-register divisions, lexical elements involved in grammatical constructions (cf. the particles above) or any others).

Given the potential wealth of results from simultaneous comparisons, this aspect of these methods should be especially welcome to researchers involved in, and underscoring the need of studying, register differences. It is of course possible, however, that a rigorous bottom-up identification will show that some of the traditional distinctions do not correspond to the levels of granularity where most variability can be found. For example, Gries (forthcoming) shows that the speaking vs. writing distinction Newman and Rice argue for in recent work (see Newman and Rice, forthcoming) yields quantitatively different results, but not qualitatively different theoretical conclusions. Similarly, a considerable body of work conducted in the QLVL Research Group at KU Leuven has been concerned with documenting lectal variation such as, for example, the differences between Belgian and Netherlandic Dutch. However, this approach is really just an instance of the kind of (well-educated) guessing at which level of granularity (variety) and between which levels of this level (Dutch vs. Belgian Dutch) effects are to be found. True, many results showed that this distinction sometimes does explain a considerable amount of variance, but a more exhaustive bottom-up approach may well show that dividing the corpus data differently may reveal factors with a higher degree of explanatory power or effect size and/or theoretical relevance (see Gries, forthcoming, for discussion).

Fourthly, the approach proposed here is based on effect sizes and, thus, avoids the shortcomings of null-hypothesis significance testing (e.g., the-sample-size-is-always-big-enough problem). However, it can also be used to approximate *p*-values without making/violating any distributional assumptions. Finally, it allows us to assess the tricky issue of corpus homogeneity with respect to two parameters that are usually not even motivated explicitly: the degree of granularity, which is almost exclusively implemented by simply taking files, and the parameter of interest, which is usually implemented by using word frequencies. The method proposed here allows us to use any degree of granularity and one's parameter of interest. Also, it is possible to both identify the parts of one's corpus which are most/least representative of the whole corpus as well as quantify the homogeneity of one's own corpus for comparison with other corpora.

Once again, note that in spite of all the benefits the proposed methods actually do not require much effort. As pointed out above, the data used here are usually available anyway. For example, retrieving all instances of the present perfect typically entails that the concordance comes with the information of which file each instance comes from, so that the classification into (sub-)registers is already done; the same holds for the other corpus data used in this paper. Also, no software other than R is necessary to do the retrieval as well as all computations and graphics. Given that so much can be achieved with so little, I hope the approach presented here, or at least the logic underlying it, offers some food for thought and will be used to improve the descriptive accuracy and explanatory understanding of variability in a wide range of corpus-based studies.

References

Benedetto, D., E. Caglioti and V. Loreto. 2002. 'Language trees and zipping', Physical Review Letters 88, 048702.

Berglund, Y. 1997. 'Future in present-day English: corpus-based evidence on the rivalry of expressions', ICAME Journal 21, pp. 7–19.

Biber, D. 1990. 'Methodological issues regarding corpus-based analyses of linguistic variation', Literary and Linguistic Computing 5 (4), pp. 257–69.

Biber, D. 1993. 'Representativeness in corpus design', Literary and Linguistic Computing 8 (4), pp. 243–57.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow, Essex: Pearson Education.

Bock, J.K. 1986. 'Syntactic persistence in language production', Cognitive Psychology 18, pp. 355–87.

Church, K.W. 2000. 'Empirical estimates of adaptation: the chance of two *Noriega*s is closer to $^{p}/_{2}$ than p^{2}', Proceedings of the Eighteenth International Conference on Computational Linguistics, pp. 180–86.

Crawley, M.J. 2002. Statistical Computing: An Introduction to Data Analysis Using S-Plus. Second printing with corrections (2004). Chichester: John Wiley and Sons.

Denoual, E. 2006. ‘A method to quantify corpus similarity and its application to quantifying the degree of literality in a document’, International Journal of Technology and Human Interaction 2 (11), pp. 51–66.

DeRoeck, A., A. Sarkar and P.H. Garthwaite. 2004. ‘Defeating the homogeneity assumption: some findings on the distribution of very frequent terms’, Technical Report 2004/7, The Open University.

Dubois, B.L. 1972. The meaning and the distribution of the perfects in present-day American English writing. Unpublished Ph.D. dissertation, University of New Mexico.

Egan, T. 2002. ‘*Let*, *allow*, *stop* and *cease*: a corpus study’, Paper presented at the International Conference on Construction Grammar. Helsinki, Finland, 6–8 September 2002.

Elsness, J. 1997. The Perfect and the Preterite in Contemporary and Earlier English. Berlin, New York: Mouton de Gruyter.

Evert, S. and B. Krenn. 2005. ‘Using small random samples for the manual evaluation of statistical association measures’, Computer Speech and Language 19 (4), pp. 450–66.

Facchinetti, R. 2001. ‘*Can* and *could* in contemporary British English: a study of the ICE-GB Corpus’ in P. Peters, P. Collins and A. Smith (eds.) New Frontiers of Corpus Research, Proceedings from the 21st International Conference on English Language on Computerized Corpora, pp. 229–46. Amsterdam: Rodopi.

Goldberg, A.E. 1995. Constructions: A Construction Grammar Approach to Argument Structure. Chicago, IL: The University of Chicago Press.

Gries, St. Th. 2003. Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement. London, New York: Continuum Press.

Gries, St. Th. 2005a. ‘Syntactic priming: a corpus-based approach’, Journal of Psycholinguistic Research 34 (4), pp. 365–99.

Gries, St. Th. 2005b. ‘Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff’, Corpus Linguistics and Linguistic Theory 1 (2), pp. 277–94.

Gries, St. Th. Forthcoming. Corpus data in usage-based linguistics: What’s the right degree of granularity for the analysis of argument structure constructions?

Gries, St. Th. and A. Stefanowitsch. 2004. ‘Extending collostructional analysis: a corpus-based perspectives on “alternations”’, International Journal of Corpus Linguistics 9 (1), pp. 97–129.

Gries, St. Th., B. Hampe and D. Schönefeld. Forthcoming. ‘Converging evidence II: More on the association of verbs and constructions’ in J. Newman and S. Rice (eds.) Empirical and Experimental Methods in Cognitive/Functional Research. Stanford, CA: CSLI.

Herzlík, B. 1976. 'Some notes on the Present Perfect', Brno Studies in English 12, pp. 57–83.

Hofland, K. and S. Johansson. 1982. Word frequencies in British and American English. Bergen: Norwegian Computing Centre for the Humanities / London: Longman.

Johansson, C. 2001. 'Pied piping and stranding from a diachronic perspective' in P. Peters, P. Collins and A. Smith (eds.) New Frontiers of Corpus Research. Proceedings from the 21st International Conference on English Language on Computerized Corpora, pp. 147–62. Amsterdam: Rodopi.

Johansson, S. and K. Hofland (eds.). 1989. Frequency Analysis of English Vocabulary and Grammar, Based on the LOB Corpus. Oxford: Clarendon Press.

Kaufman, L. and P.J. Rousseeuw. 1990. Finding Groups in Data. New York: John Wiley.

Kennedy, G. 1991. '*Between* and *through*: the company they keep and the functions they serve' in K. Aijmer and B. Altenberg (eds.) English Corpus Linguistics, pp. 95–110. London: Longman.

Kilgarriff, A. 2001. 'Comparing corpora', International Journal of Corpus Linguistics 6 (1), pp. 1–37.

Kilgarriff, A. 2005. 'Language is never, ever, ever, random', Corpus Linguistics and Linguistic Theory 1 (2), pp. 263–76.

Kilgarriff, A. and T. Rose. 1998. 'Measures for corpus similarity and homogeneity'. Proceedings of the Third Conference on Empirical Methods in Natural Language Processing, pp. 46–52.

Leech, G. and R. Fallon. 1992. 'Computer corpora: What do they tell us about culture?' ICAME Journal 16, pp. 29–50.

Manning, C.D. and H. Schütze. 2000. Foundations of Statistical Natural Language Processing. Second printing with corrections. Cambridge, MA: M.I.T. Press.

Mindt, D. 2000. An Empirical Grammar of the English Verb System. Berlin: Cornelsen.

Mukherjee, J. 2003. 'Corpus data in a usage-based cognitive grammar' in K. Aijmer and B. Altenberg (eds.) The Theory and Use of Corpora, pp. 85–100. Amsterdam: Rodopi

Newman, J. and S. Rice. 2006. 'Transitivity Schemas of English EAT and DRINK in the BNC' in St. Th. Gries and A. Stefanowitsch (eds.) Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis, pp. 225–60. Berlin, Heidelberg and New York: Mouton de Gruyter.

Oakes, M.P. 2003. ‘Text categorisation: automatic discrimination between US and UK English using the chi-square test and high ratio pairs’ Downloaded on 5 April 2006, from: http://www.cet.sunderland.ac.uk/IR/oakesRL2003.pdf

Pedersen, T., M. Kayaalp and R. Bruce. 1996. ‘Significant lexical relationships’, Proceedings of the 13th National Conference on AI (AAAI-96), pp. 455–60.

Peterson, B. 1970. ‘Towards understanding the “perfect” construction in spoken English’, English Teaching Forum 8, pp. 2–10.

R Development Core Team. 2005. R 2.2 – A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3–900051–07–0, http://www.R-project.org

Rayson, P.E. 2003. Matrix: a statistical method and software tool for linguistic analysis through corpus comparison. Unpublished Ph.D. dissertation, Computing Department, Lancaster University.

Rayson, P.E. and R. Garside. 2000. ‘Comparing corpora using frequency profiling’, Proceedings of the Workshop on Comparing Corpora, pp. 1–6.

Rayson, P.E., G. Leech and M. Hodges. 1997. ‘Social differentiation in the use of English vocabulary: some analysis of the conversational component of the British National Corpus’, International Journal of Corpus Linguistics 2 (1), pp. 133–52.

Rose, T.G., N.J. Haddock and R.C.F. Tucker. 1997. ‘The effects of corpus size and homogeneity on language model quality’, Proceedings of the ACL SIGDAT Workshop on Very Large Corpora, pp. 178–91.

Sahlgren, M. and J. Karlgren. 2005. ‘Counting lumps in word space: density as a measure of corpus homogeneity’, Proceedings of the Twelfth Edition of the Symposium on String Processing and Information Retrieval, pp. 151–54.

Schlüter, N. 2002. Present Perfect: Eine korpuslinguistische Analyse des englischen Perfekts mit Vermittlungsvorschlägen für den Sprachunterricht. Tübingen: Narr.

Schlüter, N. 2005. ‘How reliable are the results? Comparing corpus-bases studies of the present perfect’, Paper presented at the workshop ‘The scope and limits of corpus linguistics – empiricism in the description and analysis of English’; Free University, Berlin.

Sekine, S. 1997. ‘The domain dependence of parsing’, Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 96–102.

Stefanowitsch, A. and St. Th. Gries. 2003. ‘Collostructions: Investigating the interaction between words and constructions’, International Journal of Corpus Linguistics 8 (2), pp. 209–43.

Summers, D. 1996. ‘Computer lexicography – the importance of representativeness in relation to frequency’ in J. Thomas and M. Short (eds.) Using Corpora for Language Research, pp. 260–66. New York: Longman.

Summers, D. 1996. ‘Corpus lexicography – the importance of representativeness in relation to frequency’, Longman Language Review 3, pp. 6–9.

Szmrecsanyi, B. 2005. ‘Language users as creatures of habit: a corpus-based analysis of persistence in spoken English’, Corpus Linguistics and Linguistic Theory 1 (1), pp. 114–50.

Wikipedia contributors. 2006. Resampling (statistics). Wikipedia, The Free Encyclopedia. Date of last revision: 27 April 2006 19:49 UTC. Date accessed: 13 May 2006 20:58 UTC. http://en.wikipedia.org/w/index.php?title=Resampling_%28statistics %29&oldid=50468841

Xiao, Z. and A. McEnery. 2005. ‘Two approaches to genre analysis: three genres in Modern American English’, Journal of English Linguistics 33 (1), pp. 62–82.