Bodo Winter* and Martine Grice

Independence and generalizability in linguistics

https://doi.org/10.1515/ling-2019-0049
Received December 30, 2019; accepted July 8, 2021; published online August 30, 2021

Abstract: Quantitative studies in linguistics almost always involve data points that are related to each other, such as multiple data points from the same participant, multiple texts from the same book, author, genre, or register, or multiple languages from the same language family. Statistical procedures that fail to account for the relatedness of observations by assuming independence among units can lead to grossly misleading results if these sources of variation are ignored. As mixed effects models are increasingly used to analyze these non-independent data structures, it might appear that the problem of violating the independence assumption is solved. In this paper, we argue that it is necessary to re-open and widen the discussion about sources of variation that are being ignored, not only in statistical analyses, but also in the way studies are designed. Non-independence is not something that is "solved" by new statistical methods such as mixed models, but it is something that we continuously need to discuss as we apply new methods to an increasingly diverse range of linguistic datasets and corpora. In addition, our paper delivers something that is currently missing from statistical textbooks for linguists, which is an overview of non-independent data structures across different subfields of linguistics (corpus linguistics, typology, phonetics etc.), and how mixed models are used to deal with these structures.

Keywords: corpus statistics; experimental design; generalizability; mixed models; multilevel models; sampling

1 Introduction

Large-scale attempts to replicate existing studies in psychology and other fields have produced disheartening results (e.g., Camerer et al. 2018; Open Science

***Corresponding author: Bodo Winter**, Department of English Language and Linguistics, University of Birmingham, Frankland Building, Edgbaston, Birmingham, B15 2TT, UK,
E-mail: bodo@bodowinter.com
Martine Grice, Phonetics Laboratory, Institute of Linguistics, University of Cologne, Köln, Germany, E-mail: martine.grice@uni-koeln.de

Collaboration 2015). At the same time, these replication failures have generated fruitful discussion about methodological standards, including in linguistics (e.g., Berez-Kroeker et al. 2018; Roettger 2019; Roettger and Baer-Henney 2019; Roettger et al. 2019), and in this spirit the current special issue has brought linguistic researchers from various subfields together to discuss methodological challenges. In a recent paper, Yarkoni (2020) opened up another discussion on what he calls the "generalizability crisis". Yarkoni argues that researchers often make verbal statements that are more general than the corresponding studies allow. His argument rests on the idea that most studies neglect, in some form or another, known or theoretically plausible sources of variation. This can happen either in the design of a study, such as when only one item is sampled even though claims are predicated on a larger set of items, or it can happen at the analysis stage if heterogeneity across clusters in the data (such as items) is unaccounted for.

To illustrate the basic logic of Yarkoni's (2020) argument with a linguistic example, consider the claim that younger speakers use swear words more often than older speakers (Murphy 2009). Clearly, some people swear more than others. A corpus analysis that exclusively focuses on aggregate results without taking variation across speakers into account can lead to grossly misleading results (Brezina and Meyerhoff 2014; Gries 2015a; Johnson 2009; Sönning and Krug 2021; Tagliamonte and Baayen 2012). For example, the association between age and swear word usage could be driven by a few young individuals who swear a lot. Alternatively, it could be that only a specific swear word is used more often by younger speakers, not a wide range of swear word types. In both of these cases, the general statement "younger speakers use more swear words" would be incorrect as it fails to take variation across speakers and variation across words into account. Omitting important sources of variation from one's analysis limits the generalizability of any conclusions drawn from the study. Such an omission can also lead to a failure to replicate when, for example, a result was contingent on the particular item(s) chosen for a study and does not hold for an alternative item sample (Judd et al. 2012; Yarkoni 2020).

Another way of talking about the omission of variance components in one's analysis is in the context of the "independence assumption" of a statistical procedure. Independence means that for any two observations in a sample, knowing the value of one observation, relative to the mean of the population, gives us no information about the other (Kenny and Judd 1986: 422). When there is more than one observation from the same grouping unit (e.g., participants, items, texts, registers, languages, language families), these observations share a connection and cannot be assumed to be independent anymore. Failing to account for these clusters statistically violates the independence assumption.

There is an extensive literature across a diverse range of disciplines documenting the serious consequences of violating the independence assumption (e.g., Bromham et al. 2018; Hurlbert 1984; Kenny and Judd 1986; Kroodsma et al. 2001; Lazic 2010; Lazic et al. 2018, 2020; Lombardi and Hurlbert 1996; Machlis et al. 1985; Scariano and Davenport 1987; Vul et al. 2009). One of the most serious consequences is a potentially drastic increase in the rate of spuriously significant results ("Type I errors"). This has been demonstrated analytically, via simulations, and via concrete examples of published findings that cease to be significant once an analysis appropriately deals with all sources of variation that introduce non-independence (e.g., Bromham et al. 2018; Judd et al. 2012; Kenny and Judd 1986; Roberts et al. 2015; Scariano and Davenport 1987; Winter 2011).

Linguists from different subfields are already well-accustomed to avoiding violations of independence with respect to participants and items. For example, it is widely known that participants are a source of variation that needs to be incorporated into one's analysis (Baayen et al. 2008; Brezina and Meyerhoff 2014; Gradoville 2019; Gries 2015b; Johnson 2009; Tagliamonte and Baayen 2012). Likewise, the issue of bringing item variation into one's analysis has been known for a long time (Brunswik 1955; Clark 1973; Coleman 1964) and continues to be discussed in psychology (Judd et al. 2012; Wells and Windschitl 1999; Westfall et al. 2014; Yarkoni 2020). In recent years, mixed models have taken linguistics by storm precisely because they allow researchers to deal with multiple sources of non-independence in a given set of data within an integrated framework via the inclusion of random effects. Arguments for mixed models have been made in psycholinguistics (Baayen et al. 2008), sociolinguistics (Johnson 2009; Tagliamonte and Baayen 2012), typology (Jaeger et al. 2011), second language acquisition research (Cunnings 2012), and corpus linguistics (Gries 2015b).

However, mixed models do not "solve" issues of non-independence automatically (Hurlbert 2009; Yarkoni 2020). The field-specific debates about violations of independence have found a modern parallel in discussions about what happens when important random effects terms are omitted from a mixed model analysis (Aarts et al. 2015; Barr et al. 2013; Schielzeth and Forstmeier 2008; Yarkoni 2020). While there has been a debate about appropriate random effects structures in some subfields of linguistics, such as psycholinguistics (Barr et al. 2013; Matuschek et al. 2017) and typology (Jaeger et al. 2011), we were inspired by Yarkoni (2020) to broaden the focus. What sources of non-independence are there across different subfields of linguistics? And how are mixed models used across different subfields of linguistics to deal with these non-independences? Finally, are there sources of variation that are left unaccounted for, either in study design or in statistical analysis? If so, what are these sources of variation?

We think that it is crucial to re-open the discussion about the independence assumption for several reasons. First, while issues of non-independence have extensively been discussed in academic journals, pedagogical texts for linguists lag behind. For example, several new textbooks targeted primarily at corpus linguists (Desagulier 2017; Stefanowitsch 2020; Wallis 2021) do not present much discussion of the independence assumption, even though this has been the focus of extensive debate in corpus linguistics (Baroni and Evert 2009; Brezina and Meyerhoff 2014; Evert 2006; Gradoville 2019; Gries 2015a, 2015b, 2018; Kilgarriff 1996, 2005; Koplenig 2019; Lijffijt et al. 2016; Oakes and Farrow 2006).[1] For example, Desagulier (2017), Stefanowitsch (2020), and Wallis (2021) discuss the application of Chi-square tests to corpus data even though this procedure assumes independent observations (Gries 2015a, 2015b; Lijffijt et al. 2016). As pointed out by Gries (2015b: 121), mixed models could "end the way in which corpus linguists nearly always violate basic assumptions of our statistical tests". Unfortunately, mixed models are sometimes not discussed at all (e.g., Desagulier 2017; Wallis 2021) despite persistent calls for their utility in corpus linguistics (Gries 2015b; Tagliamonte and Baayen 2012). Given that corpora always have complex nested data structures that standard significance tests cannot account for, we should be asking ourselves the question whether at this stage in the methodological development of linguistics, it is at all appropriate to teach classical significance tests (cf. Koplenig 2019) instead of model-based approaches that allow us to deal with multiple sources of variation more effectively.

A second reason to re-open the discussion surrounding non-independence has to do with the introduction of new methods to linguistics. Consider, for example, random forests (Breiman 2001), a relatively new technique argued to perform well in "low n, high p" situations (Strobl et al. 2009), involving potentially small datasets (n) with a large number of potentially collinear predictors (p) (as is the case when a large number of predictors are highly correlated). Gries (2019) rightly draws attention to important methodological issues that may arise in the application of random forests to linguistic data. However, a much more fundamental issue not discussed by Gries (2019) is whether the random forest applications commonly used by linguists (such as the R packages *ranger, randomForest*, and *party*) can actually be applied to linguistic datasets at all, given that, just like other statistical procedures, random forests are biased when the data contains non-

1 It should be pointed out that not all of these texts discuss the "independence assumption" with this exact terminology, such as Evert (2006) and Kilgarriff (2005), who frame their discussion in terms of "randomness". However, in some form or another, each one of these texts deal with the independence assumption, and many reference it directly (e.g., Gries 2015a, 2015b; Kilgarriff 1996; Lijffijt et al. 2016).

independent subgroups (Hajjem et al. 2014; Karpievitch et al. 2009; Stephan et al. 2015). This shows that we need to continue the discourse of non-independence as new methods enter our field.

Third and finally, Yarkoni (2020) asks us to think about unmeasured sources of variation more widely. When it comes to study design, it is important to discuss whether particular experimental designs underestimate the variation that is present in natural language. Alternatively, we may ask ourselves whether particular experimental design strategies that have become traditions in some fields actually capture relevant sources of variation that are of interest.

Our discussion proceeds as follows. First, we use mixed models to elucidate the problem of non-independent data structures, thereby also demonstrating to readers unfamiliar with this method how mixed models allow addressing violations of the independence assumption (Section 2). Second, we give an overview of how the issue of non-independence arises in different forms across different subfields (Section 3), also discussing less commonly considered non-independence issues, as well as how mixed models can or are being used to address them.

2 Introduction to mixed models, independence, and generalizability

This section uses mixed models to elucidate the problem of omitting important sources of variation that are present in the data. It also serves as a brief primer for readers unfamiliar with mixed models that will help with the discussion that follows. Figure 1(a) shows 28 data points in which a response (y) differs as a function of a predictor (x). This scenario mimics many common situations in linguistic data analysis. For example, if this were psycholinguistic data, the predictor on the x-axis could be word frequency and the response on the y-axis could be response time: more frequent words are processed faster, as indicated by the negative slope of the superimposed linear model (thick black line).

The simple linear model in Figure 1(a) can be expressed in the following form:

1. $y_i \sim \text{Normal}\ (\mu_i, \sigma)$
2. $\mu_i = \alpha + \beta x_i$

The dependent variable, y, is assumed to follow a normal distribution with a specified mean μ ("mu") and a standard deviation σ ("sigma"). The subindex i represents the fact that this model predicts different means for different data points ($i = 1$, $i = 2$ etc.). In our model, the mean, or expected value, for observation i (i.e., μ_i) depends on the value of the predictor x (i.e., x_i). The mean itself is a function of a linear combination of an intercept (α) and a slope (β), corresponding to the two

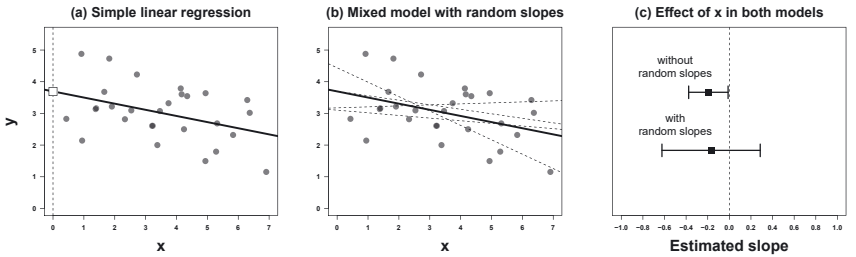

Figure 1: (a) A simple linear model shows a declining trend; the white square represents the intercept; (b) the same plot with superimposed random effects predictions for individuals; (c) the 95% credible intervals for the corresponding slopes of a simple linear model (no random effects) and a mixed model with a random effect for subject.

terms that form the equation of a line. After having seen the data, the linear model estimates the intercept to be $\widehat{\alpha} = 3.7$ and the slope to be $\widehat{\beta} = -0.2$ (the hat indicates that these are now estimates derived from the data). Here, the slope of −0.2 tells us that responses to frequent words are faster. For an extensive book-length treatment of linear models in linguistics, see Winter (2019).

A caveat with the above linear model is that it assumes that all data points are independent, an assumption shared with almost all statistical procedures commonly used in linguistics. In an experimental context, this assumption would be satisfied if each of the 28 data points came from a different participant. Let us now assume that the data seen in Figure 1(a) does, in fact, come from only a few individuals, four to be exact, as represented by the four dashed lines in Figure 1(b). This immediately changes our understanding of what "*N*" is in this data. There are not in fact 28 independent data points, but when focusing on participants, *N* is merely 4. Not incorporating information about the individual participant into the model amounts to artificially increasing sample size, which leads to a spurious increase in statistical power (Hurlbert 1984, 2009).[2] Thus, we need to bring the individual into the analysis, which can be done via a mixed model. One way of representing the mixed model corresponding to Figure 1(b) mathematically is as follows.

1. $y_i \sim \text{Normal}\,(\mu_i, \sigma)$
2. $\mu_i = \alpha_j + \beta_j x_i$
   a. $\alpha_j \sim \text{Normal}\,(0, \sigma_a)$
   b. $\beta_j \sim \text{Normal}\,(0, \sigma_\beta)$

2 Statistical power is the probability that a hypothesis test correctly rejects the null hypothesis (i.e., the probability of obtaining a significant result when the null hypothesis is actually false).

The new indented terms correspond to the random intercepts (α_j) and random slopes (β_j). Notice that the equation of a line $\mu_i = \alpha_j + \beta_j x_i$, now has additional subindices for the intercept and slope, with j representing different individuals (participant $j = 1$, participant $j = 2$, etc.).

The effects that the incorporation of random effects has on our inferences are shown in Figure 1(c), where the 95% interval of the slope estimate widens drastically (Aarts et al. 2015; Barr et al. 2013; Schielzeth and Forstmeier 2008; Yarkoni 2020).[3] The certainty that is suggested by the narrower interval of the linear model without random effects is spurious, purely resulting from the neglect of variation across individuals.

To the reader unfamiliar with mixed effects models, it should be pointed out that all the arguments presented here conceptually also apply to all standard significance tests, most of which assume independence in some form or another. Mixed models are a *solution* to the independence issues that arise when using standard significance tests, but only if all important sources of variation have actually been incorporated into the model as random effects. It is also important to emphasize that the above arguments are couched in terms of continuous data for the sake of exposition but carry over to categorical data structures and tests for categorical data (such as Chi-square tests, Fisher's exact tests etc.). Finally, to be extra clear it has to be emphasized that while the example only used "individual" (e.g., participant) to demonstrate conceptual matters, the same principles carry over to any other grouping factor (e.g., items, registers, different texts in a corpus etc.).

3 Overview of non-independent data structures in linguistics

3.1 Introduction

The literature on violations of the independence assumption across different fields, such as ecology (e.g., Hurlbert 1984), animal behavior research (e.g., Kroodsma 1989; Lombardi and Hurlbert 1996), and psychology (Kenny and Judd

3 All intervals reported here are 95% Bayesian credible intervals rather than confidence intervals (Morey et al. 2016). This does not, however, matter for our arguments. A 95% credible interval can be interpreted as having a 95% probability of containing the true value. The code used to produce this analysis is available at https://osf.io/zdrpc/.

1986) is testament to the fact that this is a serious issue that takes many different forms (Hurlbert 2009). Kenny and Judd (1986) distinguish between "non-independence due to groups", "non-independence due to sequence", and "non-independence due to space". As mentioned above, participant- and item-specific variation has already received extensive discussion in linguistics and will therefore not be repeated here. Both of these factors relate to Kenny and Judd's "non-independence due to groups", with either participant or item being the corresponding grouping factor. However, as we will outline in this section, there are other grouping factors to consider that are less commonly discussed as random effects.

The following section provides a tour of selected domains that demonstrate various forms of non-independence in linguistics, corresponding to different sources of variation that affect different kinds of linguistic data. We consider spatial dependence (Section 3.1), language and language family dependence (Section 3.2), temporal and sequence dependence (Section 3.3), talker effects (Section 3.4), dyad effects (Section 3.5), exact repetitions (Section 3.6), and the nested hierarchical structure of corpora (Section 3.7). While some of these sources of variation have received extensive discussion, others have not. It should be noted that while most of the dependencies we discuss here can be dealt with via random effects in a mixed model context, some of the dependencies (such as temporal and sequence dependencies) are dealt with in other ways, and yet other dependencies are more focused on issues relating directly to study design.

3.2 Spatial dependence

All else being equal, individuals closer to each other are more similar to each other. A linguistic reflection of this fact is that speakers closer to each other are more likely to share linguistic features, thus introducing a form of non-independence when a study includes multiple speakers from the same group. One form of dealing with this is to have a "dialect" random effect,[4] or "location" or "province" random effects (Bischetti et al. 2021; Wieling et al. 2011, 2014). Spatial dependence problems also arise in typology (Bickel 2011; Cysouw 2010; Gast and Koptjevskaja-Tamm 2018) and in studies of cultural and language evolution more generally

4 In many studies of specific dialect comparisons, "dialect" is a fixed effect (e.g., McCloy et al. 2015). This is appropriate when generalizability claims are predicated on a specific set of dialects that are of interest to a researcher. If, however, the goal is to make generalizations *across* dialects, "dialect" should be a random effect.

(Bromham et al. 2018; Roberts and Winters 2013; Roberts et al. 2015). Languages and cultures that are geographically closer to each other are more similar to each other, either because they have on average more contact (Jaeger et al. 2011; Roberts et al. 2015), or because their environments share more features (see discussion in Bromham et al. 2018).

Within a mixed model framework, this source of variation can be dealt with by adding area-based random effects, which is frequently done in typological research (e.g., Bentz and Winter 2014; Jaeger et al. 2011; Sóskuthy and Roettger 2020). There is room for discussion about what type of area-based random effects structure are appropriate, and how granular the spatial resolution should be. For example, Sóskuthy and Roettger (2020) use macro-areas from Glottolog (Hammarström et al. 2020), which divide the world into six areas (Africa, Australia, Eurasia, North America, Papunesia, and South America). While this accounts for the fact that languages within each area are expected to be more similar to each other due to contact, it does not account for more fine-grained within-area structures (e.g., within Africa, two languages within the same country or region are expected to be more similar to each other, see Bromham et al. 2018).

3.3 Language or language family

Just as "dialect" is a viable random effect, "language" can be a viable random effect for crosslinguistic studies that have multiple data points from the same language and that seek to make generalizations over languages. An example of this is Murakami (2016), who performed a corpus analysis of L2 learners' writings with writers from 10 different languages, with "language" as random effect. If an L2 learner study is specifically interested in comparing a small set of languages, such as Chinese learners versus Korean learners of English, then it is appropriate to fit "language" as a fixed effect. However, doing so effectively constrains one's generalizations to the specific languages investigated. By not including language as a random effect, those studies cannot make conclusions about languages beyond those considered in the sample (cf. Yarkoni 2020). Thus, if the goal of an analysis is to generalize across languages, as was the case in Murakami (2016), language should be a random effect.

It is a widely discussed problem that crosslinguistic analyses need to control for genealogical dependencies (Bickel 2011; Bromham et al. 2018; Cysouw 2010). As a result of this, typological studies generally include language family as a random effect alongside the above-mentioned area effects (e.g., Bentz and Winter 2014; Jaeger et al. 2011; Sóskuthy and Roettger 2020). As an example of what can happen when genealogical dependencies are ignored, consider Chen (2013), who reported

that speakers of languages without a grammaticized present/future distinction save more money than speakers of languages who do make this grammatical distinction. Roberts et al. (2015) showed that this association between grammar and savings behavior goes away in a mixed model analysis with random effects for language family and area. In a similar case, Atkinson (2011) suggested that languages situated further away from Africa have fewer phonemes, a result which also did not survive statistical controls for genealogical and areal dependencies (Jaeger et al. 2011). Bromham et al. (2018) discuss some of the caveats with having language family as a random effect: while this accounts for the fact that two or more languages within the same family are more similar to each other, it does not deal with the more fine-grained aspects of a phylogeny. For example, within Indo-European, German and English are more closely related to each other than German and French. Bromham et al. (2018) discuss additional approaches to control for phylogenetic dependencies.

3.4 Temporal and sequence dependence

Serial dependence is a common feature for sequential or time series data, where observations are often more similar to each other as a function of time lag (autocorrelation). Serial dependency comes up in both experimental and corpus linguistic contexts. First, with respect to experiments, any study that involves repeated measures from the same individual automatically introduces time or sequence as a dependency factor. This includes such things as fatigue, attentional fluctuations, or learning over successive trials. Baayen et al. (2017) show that psycholinguistic data contains a non-negligible amount of such serial dependency that also interacts with condition predictors. Sequence dependence is not the sort of idiosyncratic variation that is dealt with via random effects and therefore requires a different analysis approach. Baayen et al. (2017) show that Generalized Additive Mixed Models (GAMMs) can be used to deal with this (for tutorials, see Sóskuthy 2017; Wieling 2018; Winter and Wieling 2016). GAMMs are an extension of mixed models that can be used to model time series data via the incorporation of nonlinear transformations of predictors, such as "time" or "trial order."

Baayen et al.'s (2017) arguments also extend to other areas of linguistics, such as phonetics. For example, in a speech production experiment, speakers could be more or less prone to hyperarticulation over successive trials. Such trial-order-based dependencies are not commonly considered in the statistical analysis of phonetic data. Finally, serial dependence also arises in corpus linguistics, as has been discussed extensively in the literature on syntactic priming or structural persistence (e.g., Gradoville 2019; Gries 2005; Sankoff and Laberge 1978;

Szmrecsanyi 2005). In this case, a speaker or writer is more likely to use a particular construction again after it has recently been used, thus making expressions syntagmatically closer to each other more statistically dependent. Kilgarriff (1996) describes this as "clumpiness" and speaking of lexical data, he says that "words come in clumps; unlike lightning, they often strike twice" (p. 4).

3.5 Talker effects

Many psycholinguistic or phonetic experiments involve stimuli that are created by a single model speaker. We know that which speaker one listens to influences speech processing (e.g., Buchan et al. 2008; Creel and Bregman 2011; Hay et al. 2009; Trude and Brown-Schmidt 2012), as also evidenced by the literature on talker effects in cross-dialect and accent perception (e.g., Flege and Fletcher 1992; McCloy et al. 2015). One way of dealing with this statistically is by adding "talker" as a random effect to a mixed effects analysis. McCloy et al.'s (2015) mixed model analysis of cross-dialect speech intelligibility actually found more random effects variance for by-talker differences than for by-listener differences.

The fact that listeners adapt to a particular talker's idiosyncratic speech features (Nygaard and Pisoni 1998; Nygaard et al. 1994) has wide-ranging implications for speech perception research. The same way that we can think of sampling items, as is common in psycholinguistic research, we should consider the option of sampling voices as well. If only one talker produces all stimuli, this effectively constrains any generalizations to the model speaker used, a problem that has also been raised in other fields (Kroodsma 1989). In analogy to the "language-as-fixed-effect fallacy" (Clark 1973), this could be called the "talker-as-fixed effect fallacy." In the spirit of Yarkoni's (2020: 19) recommendation to "design with variation in mind", we have conducted perception experiments which sample different voices, and in which "voice" was added as a random effect (Baumann and Winter 2018; Brown et al. 2014; Cangemi et al. 2015; Idemaru et al. 2020; Roettger et al. 2014). This practice has also been adopted by some researchers conducting speech perception studies in sociolinguistics (e.g., Ruch 2018).

3.6 Dyad effects

Dyadic communication is fundamentally different from linguistic tasks that involve just one person, and a statistical analysis of dyadic data needs to reflect this. Peters et al. (2014) conducted a carefully constructed study on the phonetic realization of focus in several Germanic languages. To collect naturalistic data,

participants completed short mini-dialogs in pairs. After completing the task once, participants switched roles, a practice that is common when pairs of subjects, rather than a subject and a confederate, enact a dialog. This is the standard procedure when recording task oriented dialogs such as the widely employed Map Task (Anderson et al. 1991), in which participants draw routes on multiple maps, switching between instruction-giver and instruction-follower for a second task with the same partner.

We know from much phonetic research on what is called "accommodation" or "convergence" that when interacting with others, there is a strong tendency for speaker's utterances to become more similar to the interlocutor (e.g., Abel and Babel 2017; Giles and Powesland 1997; Nielsen 2011; Pardo 2006). The existence of accommodation adds a layer of statistical dependence to any analysis that includes data from two or more speakers that talk to each other during experimental tasks. When participants in the study of Peters and colleagues (2014) respond with a focus type, a statistical model has to take into account the fact that they have already spoken with their interlocutor, and that their focus realization may be influenced by what the interlocutor has said previously. However, even if they had used the same confederate in every dyad, the confederate could still have been influenced by productions of previous discourse partners, a possibility that would need to be taken into account in the analysis too. The problem of dyadic data structures has been extensively discussed in psychology (e.g., Kenny 1996), with one common solution being the addition of "dyad" random effects (cf. Wendorf 2002). We know of no linguistic research that used a dyadic task and modeled this with random effects.

It is worth pointing out that specific designs can introduce dyadic structures "under the hood". Cangemi et al. (2015) is a perception study that includes multiple listeners as well as multiple talkers. The fact that both listeners and talkers are repeated in this study also means that listener-talker combinations are repeated. This was dealt with by including a "dyad" random effect. Model comparison revealed that the "dyad" effect captured non-negligible variation, demonstrating that different listeners respond differently to specific talkers. Thus, even if the experimental design does not involve people directly talking to each other, there may be hidden "dyadic" structures that may be important to incorporate into one's statistical analysis.

3.7 Exact repetitions

There are clear subfield-specific traditions in the design of experiments. Many speech production experiments traditionally include what we call "exact

repetitions", which involves the same speaker producing the same item in the same linguistic context multiple times. Roettger and Gordon (2017) survey production studies of word stress, and find that 57 out of 113 studies included two or more exact repetitions; Nicenboim et al. (2018) find that 10 out of 24 studies on the phenomenon of incomplete neutralization included exact repetitions; and Winter (2015) found that 26 out of 35 experimental studies in the 2014 issue of *Journal of Phonetics* included exact repetitions (https://osf.io/zdrpc/). A quick search for "repeat" or "repetition" in the most recent issues of *Language and Speech* and *Journal of Phonetics* (all 2019 issues) reveals that this practice is still common, with 21 out of 84 papers including exact repetitions. Importantly, there seems to be no standard for dealing with repetition as a source of variation. Repetitions are variously included as fixed effect (e.g., Oh and Byrd 2019), random effect (e.g., Lee and Jongman 2019), or averaged out of the data (e.g., Chan and Hall 2019). What is perhaps most worrisome is that several studies in our most recent 2019 sample of phonetics papers do not explicitly comment on how repetitions were dealt with analytically.[5]

The question arises over what source of variation repetitions are intended to generalize. It seems that speech production researchers view repetitions as random fluctuations around a "target value", with the goal of generalizing over these fluctuations. Cho et al. (2014: 134) explain that for their study "the data were averaged over repetitions across items with different stops in order to provide each speaker's representative value per condition". Similarly, Broad and Clermont (2014: 54) explicitly comment: "For statistical stability, the data used here are the averages of these measurements over the five repetitions". In his introduction to phonetic analysis, Harrington (2010: 11) says: "since a single production of a target word could just happen to be a statistical aberration, researchers in experimental phonetics usually have subjects produce exactly the same materials many times over". Interestingly, all of these arguments could just as well be ported to other fields of linguistics, as variation is an intrinsic component of any aspect of language — yet, exact repetitions are not often considered in the design of psycholinguistic experiments, unless repetition priming itself is of theoretical interest.

What is perhaps worrisome is that in the above-mentioned sample of journal papers, item numbers are inversely correlated with the number of exact repetitions across studies. This is the case for Winter's (2015) survey (Spearman's *rho* = –0.54), as well as, albeit less so, for Roettger and Gordon (2017) (*rho* = –0.35). It seems that some phonetic researchers may trade generalization across items with generalization across exact repetitions, a practice that is putatively limiting the

5 This includes work by the authors themselves, who in a recent paper that included exact repetitions did not specify directly how they were entered into the analysis.

generalizability of a study given that items are such an important source of variation in experimental research (Baayen et al. 2008; Clark 1973; Judd et al. 2012).[6]

It is important to open up discussion about whether the logic of generalizing over repetitions to measure the "representative value" is actually sound (Winter 2015). Exact repetitions would only help us ascertain a better estimate of the true production target if the repetitions randomly vary around the value of interest, which is demonstrably not the case (e.g., Kello et al. 2008), especially due to the serial dependence effects mention above. Specifically, it is known that repeating a word leads to repetition priming or reduction (Aylett and Turk 2004; Fowler 1988; Gregory et al. 1999; Pluymaekers et al. 2005), and averaging over repetition means that these systematic repetition effects are confounded with one's estimate of the "true" value (Winter 2015). In addition, including exact repetitions makes experiments longer and less ecologically valid (Niebuhr and Michaud 2015; Winter 2015). Finally, experiments with exact repetitions introduce a new dimension of researcher degrees of freedom (Roettger 2019), given that there are many different ways of dealing with exact repetitions analytically, and currently no widely agreed-upon standards within phonetic research. For all of these reasons, we suggest that exact repetitions should never be included as an unquestioned *default* in phonetic studies unless variation over repetitions is itself of theoretical interest. And, if there is the option to increase statistical power via increasing the number of items and the number of participants, this is preferred over increasing power via exact repetitions.

3.8 Corpus linguistic data

Although it has not always been named as such (see Footnote 1, above), the problem of non-independent data structures has been extensively discussed in corpus linguistics (Baroni and Evert 2009; Brezina and Meyerhoff 2014; Evert 2006; Gradoville 2019; Gries 2015a, 2015b, 2018; Kilgarriff 1996, 2005; Koplenig 2019; Lijffijt et al. 2016; Oakes and Farrow 2006) and therefore also requires a more extensive treatment here. Even though "words within a text are not independent" (Lijffijt et al. 2016: 374), researchers have only very recently considered mixed models as a solution (Gries 2015b). In contrast to psycholinguistics, there are still relatively few studies that use mixed models. While corpus linguistics itself is a

6 Many phonetic studies may have small item numbers because constraints on the form of the words (e.g., all voiced segments for intonation research, or words with a particular type of stop in the onset) limits the set of words that can be used. However, it may be possible to include a much larger set of items that are less controlled by dealing with the different item characteristics statistically.

diverse field with subfields that have different methodological traditions, it is insightful to consider Paquot and Plonksky's (2017) review of learner corpus research, which finds that out of a sample of 378 studies, the overwhelming majority (86%) uses simple statistical tests (*t*-tests, ANOVAs, log-likelihood tests, correlation tests etc.), even though these tests cannot accommodate the presence of multiple non-independent grouping structures that are inherent to any corpus.[7]

To exemplify the non-independence issues that arise in corpus linguistics, we will adapt the logical structure of an example from Brezina and Meyerhoff (2014), who raise the important issue of speaker-specific dependencies not being accounted for in aggregate analyses of corpora (Gries 2015a, 2015b; Sönning and Krug 2021; Tagliamonte and Baayen 2012). Consider the hypothetical data shown in Table 1, based on corpus data coming from six speakers, only one of whom produces more active sentences than passive sentences. If we aggregate this result across speakers, there are 90 active sentences as opposed to 60 passive ones, a result that is significant in an exact binomial test ($p = 0.02$). As before, the choice of the test does not matter here, as the problem would equally matter for other procedures, such as those performed on contingency tables (e.g., Chi-square tests, Fisher's exact test). The online repository (https://osf.io/zdrpc/) demonstrates how the data shown in Table 1 can be analyzed with a mixed model, in which case actives are not reliably over-represented anymore, consistent with the fact that only one speaker in Table 1 exhibits this pattern.

Brezina and Meyerhoff (2014) discuss an approach that deals with speaker-specific tendencies, but mixed models are a more principled way of incorporating

Table 1: Active and passive counts for six speakers, only one of which shows an increased use of actives over passives.

Speaker	Passive	Active
1	10	40
2	10	10
3	10	10
4	10	10
5	10	10
6	10	10

7 Some tests can accommodate dependencies for one dimension at a time, such as a paired *t*-test or a random effects ANOVA. However, this will always neglect other sources of variation present in the corpus because these tests cannot accommodate multiple sources of variation at the same time, as is the case with mixed models.

multiple sources of variation, such as variation over texts, speakers, and registers — all in the same model. We need to think about the fact that for any aggregate corpus data such as the data shown in Table 1, there is not just speaker variation that needs to be taken into account, but also variation across the many other nested grouping structures that characterize corpus data, as visualized in Figure 2. Sticking to the active versus passive example, certain verbs are more to being passivized (e.g., *shortlist*, *arrest*), and passives are more frequent in certain registers (e.g., academic writing). These and other sources of variation are all ignored in any aggregate analysis. Instead, these sources of variation should be directly estimated as part of the same statistical model so that the active versus passive frequency difference can be evaluated with respect to speaker, item, and register variation simultaneously within the same model.

The widely discussed set of dispersion measures that quantify dispersion across individual texts or individual speakers (Egbert et al. 2020; Gries 2006, 2008; Gries and Ellis 2015; Tagliamonte and Pabst 2020) are testament to the fact that corpus linguistics of course has always been concerned with variation across grouping structures. Gries and Ellis (2015: 233) say that "the fact that very similar or even identical frequencies of tokens can come with very different degrees of dispersion in a corpus makes the exploration of dispersion information virtually indispensable". While dispersion measures are interesting and important in their own right (Gries and Ellis 2015), supplementing an aggregate analysis with a

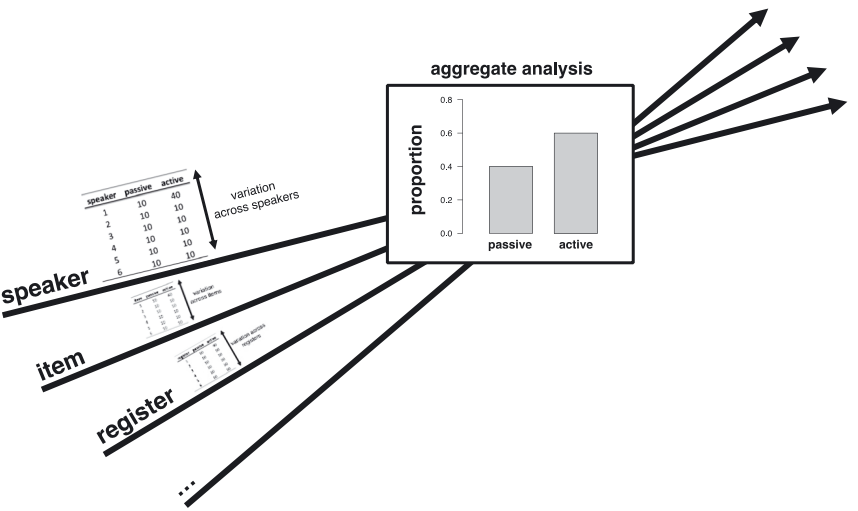

Figure 2: A corpus linguistic example where an aggregate difference (more actives than passives) simultaneously arises from multiple different sources of variation.

separate dispersion analysis means that the aggregate analysis is not formally (within the same model) evaluated against the dispersion. A unique benefit of mixed models is that they give the analyst both the aggregate result, as well as estimates of dispersion, which is reflected in the random effects variances.

As the use of mixed models in corpus linguistic research is still in its infancy, the issue of appropriate random effects structures for corpora arises. The majority of studies seem to use individual-specific random effects, such as "speaker" (Gradoville 2019; Gries 2015b; Tagliamonte and Baayen 2012; Tagliamonte and Pabst 2020), "learner" (Murakami 2016), "author" (Geleyn 2017), "tweet author" (Spina 2019), "postgraduate student" (Nasseri 2021), or "scribe" (Barteld et al. 2016).[8] In a similar fashion, several studies use mixed models to generalize over different types of texts. To look at dispersion across different texts, corpora that come in separate files allow fitting "file" as a random effect (Gries 2015a; Levshina 2016; Röthlisberger et al. 2017; Szmrecsanyi 2019; Szmrecsanyi et al. 2016), or "website" in web corpus data (Levshina 2018). Other researchers have used "register" as random effect to show that results generalize over register variation (Szmrecsanyi 2019; Szmrecsanyi et al. 2016; Wolk et al. 2013). Corpora involve complex nesting of registers, subregisters, text types, genres etc.,[9] which are exactly the types of hierarchical structures that mixed models are used for in the social sciences (Gelman and Hill 2006).

Finally, and very importantly, item-specific random effects also need to be considered, such as when showing that generalizations such as "active voice is more frequent than passive voice" generalize over different verbs (Clark 1973). In line with this, several corpus studies have included item-level variables such as "verb" as random effects into their mixed models (Bresnan et al. 2007; De Smet and Van de Velde 2020; Geleyn 2017; Gries 2015b; Grieve et al. 2019; Levshina 2016, 2018; Röthlisberger et al. 2017). Interestingly, a look at the use of mixed models in corpus linguistic research suggests that studies either include item-specific random effects or text-grouping-specific random effects (speaker, file, register etc.), but rarely both at the same time. Moreover, one is hard-pressed to find corpus linguistic studies that fit random slopes, even though it is known that the omission of random slopes can make mixed model analyses anti-conservative (Aarts et al. 2015; Barr et al. 2013; Schielzeth and Forstmeier 2008).

8 Paolillo (2013) recommends fitting "speaker" as fixed effect. While this is mathematically possible, it constrains any analysis to the particular type of speakers that are sampled, thus directly limiting the generalizability of any conclusions to the specific sample at hand. It is generally not advisable to consider "speaker" as fixed (cf. Judd et al. 2012).

9 See, for example, the multiple nested structures of the ICE International Corpus of English: https://web.archive.org/web/20200203043847/http://ice-corpora.net/ice/design.html.

It has to be acknowledged, however, that there are several practical issues with the application of mixed models to corpus linguistic data. While many corpora come in file structures that uniquely mark speakers/authors, Schäfer (2019) notes that many web corpora do not come with information about authors, and the same applies to many other corpora, especially when standard corpus query interfaces are used without access to the raw data. In some cases, it is possible to derive random effects directly from the structure of the data itself. For instance, the internal structure of a document such as sentence breaks, paragraph breaks and chapter breaks could be used to create identifiers for random effect levels. Adding sentence/paragraph/chapter random effects would help account for the fact that linguistic features are often clustered in text (Baroni and Evert 2009), as also attested by the above-mentioned literature on structural priming/persistence. A very promising approach to deal with the absence of structural metadata has been spearheaded by Pijpops et al. (2018), who used distributional semantics to group chunks of texts into semantic clusters. These bottom-up derived semantic clusters were then used as random effects levels in a mixed model analysis.

Another limiting factor is that mixed models become harder to estimate with more complex random effects structures. Schäfer (2019) discusses how his choice of random effects was constrained by which models actually converged. Convergence, however, is much facilitated when mixed models are estimated in a Bayesian framework (e.g., Nalborczyk et al. 2019; Sorensen and Vasishth 2015), which some people have begun to apply to corpus data (Levshina 2018). In fact, corpus linguistics should be in a position to fit much more complex random effects structures than is possible compared to such fields as psycholinguistics, given that there is more data and often also more random effects levels, both of which generally facilitate convergence.

4 Conclusions

In this paper, we have given an overview of non-independent data structures across subfields of linguistics and how these can be tackled via study design and statistical analysis. This overview also serves to show that non-independence, far from being an issue that is solved with the advent of mixed models in linguistics, is something that requires continued discussion and education. Yarkoni's (2020) arguments for the "generalizability crisis" are focused on psychology, and he uses examples where experiments fail to demonstrate generalization across stimuli. In many ways, the stimulus problem is something that linguistics does not have, as it is by now well-established practice to incorporate items as random effects into experimental design and statistical analysis. However, our discussion of talker

effects, dyads, exact repetitions, and other sources of variation shows that there are additional discussions to be had about which sources of variation linguists wish to generalize over, and how these sources of variation are incorporated into statistical analysis. We also need to consider how certain experimental designs undersample the variation that would be present in more naturalistic settings, such as using a single voice as stimulus in a speech perception experiment when in fact we wish to make generalizations over voices, or recording few speakers and items with many repetitions when in fact more generalizable results can be achieved with more speakers, more items, and fewer repetitions. Relieved by the constraints that ANOVAs and classical statistical tests impose on experimental designs, it is possible to design experiments that bring more of the natural heterogeneity of language back into the data, therefore allowing us to model these sources directly via mixed models. This answers Yarkoni's (2020: 19) call to "design with variation in mind".

Linguistics is uniquely positioned to become one of the sciences with the highest potential for generalizability, in large part due to the availability of large-scale corpora (Grieve this issue), which naturally harbor a lot of variability, allowing for more ecologically valid and wide-reaching generalizations. However, this can only be achieved if we actually use the tools that allow making these generalizations in statistical terms, such as mixed models. Moreover, the discussion about appropriate random effects structures in psycholinguistics (Barr et al. 2013; Matuschek et al. 2017) needs to be expanded to subfields where this discussion has not been had or taken an actual effect yet, such as corpus linguistics. Finally, we will only solve any "crisis" of generalizability (Yarkoni 2020), as well as any replication or even credibility crisis more generally, if we update our pedagogy. Unfortunately, statistical training still emphasizes classical significance tests at the expense of model-based approaches, even though for any reasonably complex data set, it is practically impossible not to violate the independence assumption with respect to some dimension of non-independence when using standard significance tests. Therefore, textbooks need to actively warn learners about "the serious consequences that result from ignoring certain variance components" (Judd et al. 2012: 55). Thus, non-independence and with it the connected topic of generalizability is something that needs continued discussion, as well as continued statistical education.

Acknowledgments: We thank Jason Grafmiller, Timo Roettger, and Akira Murakami for helpful comments and suggestions on an earlier draft of this manuscript.
Research funding: This work was supported by the German Research Foundation as part of the Collaborative Research Centre CRC-1252 "Prominence in Language"

(281511265). In addition, Bodo Winter was supported by the UKRI Future Leaders Fellowship MR/T040505/1.

References

Aarts, Emmeke, Conor V. Dolan, Matthijs Verhage & Sophie van der Sluis. 2015. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience* 16(1). 94.

Abel, Jennifer & Molly Babel. 2017. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech* 60(3). 479–502.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister & Jim Miller. 1991. The HCRC map task corpus. *Language and Speech* 34(4). 351–366.

Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027). 346–349.

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Baayen, Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Baayen, Harald, Shravan Vasishth, Reinhold Kliegl & Bates Douglas. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234.

Baroni, Marco & Stefan Evert. 2009. Statistical methods for corpus exploitation. In Lüdeling Anke & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 777–803. Berlin & New York: Mouton de Gruyter.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Barteld, Fabian, Stefan Hartmann & Renata Szczepaniak. 2016. The usage and spread of sentence-internal capitalization in early new high German: A multifactorial approach. *Folia Linguistica* 50(2). 385–412.

Baumann, Stefan & Bodo Winter. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70. 20–38.

Bentz, Christian & Bodo Winter. 2014. Languages with more second language learners tend to lose nominal case. In Søren Wichmann & Jeff Good (eds.), *Quantifying language dynamics*, 96–124. Leiden: Brill.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Heston Tyler, Gary Holton, Pulsifer Peter, David I. Beaver, Shobhana Chelliah, Dubinsky Stanley, Richard P. Meier, Nick Thieberger, Keren Rice, C Anthony & Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Bickel, Balthasar. 2011. Absolute and statistical universals. In Patrick C. Hogan (ed.), *The Cambridge encyclopedia of the language sciences*, 77–79. Cambridge: Cambridge University Press.

Bischetti, Luca, Paolo Canal & Valentina Bambini. 2021. Funny but aversive: A large-scale survey of the emotional response to Covid-19 humor in the Italian population during the lockdown. *Lingua* 249. 102963.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.

Bresnan, Joan, Cueni Anna, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Proceedings of the KNAW Academy colloquium: Cognitive foundations of interpretation*, 69–94. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen.

Brezina, Vaclav & Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1). 1–28.

Broad, David J. & Frantz Clermont. 2014. A method for analyzing the coarticulated CV and VC components of vowel-formant trajectories in CVC syllables. *Journal of Phonetics* 47. 47–80.

Bromham, Lindell, Hua Xia, Marcel Cardillo, Hilde Schneemann & Simon J Greenhill. 2018. Parasites and politics: Why cross-cultural studies must control for relatedness, proximity and covariation. *Royal Society Open Science* 5(8). 181100.

Brown, Lucien, Bodo Winter, Kaori Idemaru & Sven Grawunder. 2014. Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics* 66. 45–60.

Brunswik, Egon. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62(3). 193.

Buchan, Julie N., Martin Paré & Kevin G. Munhall. 2008. The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research* 1242. 162–171.

Camerer, Colin F., Dreber Anna, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek & Thomas Pfeiffer. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2(9). 637–644.

Cangemi, Francesco, Martina Krüger & Martine Grice. 2015. Listener-specific perception of speaker-specific production in intonation. In Susanne Fuchs, Daniel Pape, Caterina Petrone & Pascal Perrier (eds.), *Individual differences in speech production and perception*, 123–145. Frankfurt: Peter Lang.

Chan, Kit Ying & Michael D. Hall. 2019. The importance of vowel formant frequencies and proximity in vowel space to the perception of foreign accent. *Journal of Phonetics* 77. 100919.

Chen, M. Keith. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review* 103(2). 690–731.

Cho, Taehong, Yoonjeong Lee & Sahyang Kim. 2014. Prosodic strengthening on the/s/-stop cluster and the phonetic implementation of an allophonic rule in English. *Journal of Phonetics* 46. 128–146.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12(4). 335–359.

Coleman, Edmund B. 1964. Generalizing to a language population. *Psychological Reports* 14(1). 219–226.

Creel, Sarah C. & Micah R. Bregman. 2011. How talker identity relates to language processing. *Language and Linguistics Compass* 5(5). 190–204.
Cunnings, Ian. 2012. An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28(3). 369–382.
Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14(2/3). 253–286.
De Smet, Isabeau & Freek Van de Velde. 2020. A corpus-based quantitative analysis of twelve centuries of preterite and past participle morphology in Dutch. *Language Variation and Change* 32(2). 241–265.
Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Berlin: Springer.
Egbert, Jesse, Brent Burch & Biber Douglas. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115.
Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–190.
Flege, James Emil & Kathryn L. Fletcher. 1992. Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America* 91(1). 370–389.
Fowler, Carol A. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech* 31(4). 307–319.
Gast, Volker & Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology. In Daniël Van Olmen, Tanja Mortelmans & Brisard Frank (eds.), *Aspects of linguistic variation*, 43–82. Berlin & Boston: De Gruyter Mouton.
Geleyn, Tim. 2017. Syntactic variation and diachrony. The case of the Dutch dative alternation. *Corpus Linguistics and Linguistic Theory* 13(1). 65–96.
Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
Giles, Howard & Peter Powesland. 1997. Accommodation theory. In Nikolas Coupland & Adam Jaworski (eds.), *Sociolinguistics*, 232–239. Berlin: Springer.
Gradoville, Michael. 2019. The role of individual variation in variationist corpus-based studies of priming. *Italian Journal of Linguistics* 30(1). 93–124.
Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier & Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35. 151–166.
Gries, Stefan. 2006. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 191–202.
Gries, Stefan. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.
Gries, Stefan. 2015a. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93–117.
Gries, Stefan. 2015b. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.
Gries, Stefan. 2018. On over-and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1(2). 276–308.
Gries, Stefan. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3). 617–647.

Gries, Stefan & Nick C Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning* 65(S1). 228–255.
Gries, Stefan T. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.
Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami & Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2. 11.
Hajjem, Ahlem, François Bellavance & Denis Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84(6). 1313–1328.
Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2020. *glottolog/glottolog: Glottolog database 4.3*. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4061162 (accessed 31 March 2021).
Harrington, Jonathan. 2010. *Phonetic analysis of speech corpora*. Chichester: John Wiley & Sons.
Hay, Jennifer, Katie Drager & Paul Warren. 2009. Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics* 29(2). 269–285.
Hurlbert, Stuart H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54(2). 187–211.
Hurlbert, Stuart H. 2009. The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology* 123(4). 434.
Idemaru, Kaori, Bodo Winter, Lucien Brown & Grace Eunhae Oh. 2020. Loudness trumps pitch in politeness judgments: Evidence from Korean deferential speech. *Language and Speech* 63(1). 123–148.
Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319.
Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1). 359–383.
Judd, Charles M., Westfall Jacob & David A. Kenny. 2012. Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103(1). 54.
Karpievitch, Yuliya V., Elizabeth G. Hill, Anthony P. Leclerc, Alan R. Dabney & Jonas S. Almeida. 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PloS One* 4(9). e7087.
Kello, Christopher T., Gregory G. Anderson, John G. Holden & Guy C. Van Orden. 2008. The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science* 32(7). 1217–1231.
Kenny, David A. 1996. Models of non-independence in dyadic research. *Journal of Social and Personal Relationships* 13(2). 279–294.
Kenny, David A. & Charles M. Judd. 1986. Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin* 99(3). 422.
Kilgarriff, Adam. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of the AISB Workshop Language Engineering for Document Analysis and Recognition*, 33–40. Brighton: University of Sussex.
Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.
Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.

Kroodsma, Donald E. 1989. Suggested experimental designs for song playbacks. *Animal Behaviour* 37. 600–609.
Kroodsma, Donald E., Bruce E. Byers, Eben Goodale, Steven Johnson & Wan-Chun Liu. 2001. Pseudoreplication in playback experiments, revisited a decade later. *Animal Behaviour* 61. 1029–1033.
Lazic, Stanley E. 2010. The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? *BMC Neuroscience* 11(1). 5.
Lazic, Stanley E., Charlie J. Clarke-Williams & Marcus R. Munafò. 2018. What exactly is 'N' in cell culture and animal experiments? *PLoS Biology* 16(4). e2005282.
Lazic, Stanley E., Jack R. Mellor, Michael C. Ashby & Marcus R. Munafo. 2020. A Bayesian predictive approach for dealing with pseudoreplication. *Scientific Reports* 10(1). 1–10.
Lee, Hyunjung & Allard Jongman. 2019. Effects of sound change on the weighting of acoustic cues to the three-way laryngeal stop contrast in Korean: Diachronic and dialectal comparisons. *Language and Speech* 62(3). 509–530.
Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268.
Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help. *Glossa: A Journal of General Linguistics* 3(1). https://doi.org/10.5334/gjgl.294.
Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki & Heikki Mannila. 2016. Significance testing of word frequencies in corpora. *Literary and Linguistic Computing* 31(2). 374–397.
Lombardi, Celia M. & Stuart H. Hurlbert. 1996. Sunfish cognition and pseudoreplication. *Animal Behaviour* 52. 419–422.
Machlis, L., P. W. D. Dodd & J. C. Fentress. 1985. The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Zeitschrift für Tierpsychologie* 68(3). 201–214.
Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Bates Douglas. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.
McCloy, Daniel R., Richard A. Wright & Pamela E. Souza. 2015. Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech* 58(3). 371–386.
Morey, Richard D., Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee & Wagenmakers Eric-Jan. 2016. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* 23(1). 103–123.
Murakami, Akira. 2016. Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning* 66(4). 834–871.
Murphy, Bróna. 2009. 'She's a fucking ticket': The pragmatics of fuck in Irish English – an age and gender perspective. *Corpora* 4(1). 85–106.
Nalborczyk, Ladislas, Cédric Batailler, Hélène Løevenbruck, Anne Vilain & Paul-Christian Bürkner. 2019. An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research* 62(5). 1225–1242.
Nasseri, Maryam. 2021. Is postgraduate English academic writing more clausal or phrasal? Syntactic complexification at the crossroads of genre, proficiency, and statistical modelling. *Journal of English for Academic Purposes* 49. 100940.

Nicenboim, Bruno, Timo Roettger & Shravan Vasishth. 2018. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70. 39–55.
Niebuhr, Oliver & Alexis Michaud. 2015. Speech data acquisition: The underestimated challenge. *Kiehler Arbeiten in Linguistik und Phonetik* 3. 1–42.
Nielsen, Kuniko. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics. Elsevier* 39(2). 132–142.
Nygaard, Lynne C. & David B. Pisoni. 1998. Talker-specific learning in speech perception. *Perception & Psychophysics* 60(3). 355–376.
Nygaard, Lynne C., Mitchell S. Sommers & David B. Pisoni. 1994. Speech perception as a talker-contingent process. *Psychological Science* 5(1). 42–46.
Oakes, Michael P. & Malcolm Farrow. 2006. Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing* 22(1). 85–99.
Oh, Miran & Dani Byrd. 2019. Syllable-internal corrective focus in Korean. *Journal of Phonetics* 77. 100933.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). aac4716.
Paolillo, John C. 2013. Individual effects in variation analysis: Model, software, and research design. *Language Variation and Change* 25(1). 89–111.
Paquot, Magali & Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1). 61–94.
Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4). 2382–2393.
Peters, Jörg, Judith Hanssen & Carlos Gussenhoven. 2014. The phonetic realization of focus in West Frisian, Low Saxon, High German, and three varieties of Dutch. *Journal of Phonetics* 46. 185–209.
Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers & Freek Van de Velde. 2018. Comparing explanations for the complexity principle: Evidence from argument realization. *Language and Cognition* 10(3). 514–543.
Pluymaekers, Mark, Mirjam Ernestus & Harald Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62(2/4). 146–159.
Roberts, Seán & James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PloS One* 8(8). e70902.
Roberts, Seán, James Winters & Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. *PloS One* 10(7). e0132145.
Roettger, Timo. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology. Journal of the Association for Laboratory Phonology* 10(1). 1.
Roettger, Timo B. & Dinah Baer-Henney. 2019. Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis* 1(4). 1–23.
Roettger, Timo & Matthew Gordon. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard* 3(1). 20170006.
Roettger, Timo, Bodo Winter, Sven Grawunder, James Kirby & Martine Grice. 2014. Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics* 43. 11–25.
Roettger, Timo B., Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.
Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4). 673–710.

Ruch, Hanna. 2018. The role of acoustic distance and sociolinguistic knowledge in dialect identification. *Frontiers in Psychology* 9. 818.
Sankoff, David & Suzanne Laberge. 1978. Statistical dependence among successive occurrences of a variable in discourse. In David Sankoff (ed.), *Linguistic variation: Models and methods*, 119–126. New York, NY: Academic Press.
Scariano, Stephen M. & James M. Davenport. 1987. The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician* 41(2). 123–129.
Schäfer, Roland. 2019. Prototype-driven alternations: The case of German weak nouns. *Corpus Linguistics and Linguistic Theory* 15(2). 383–417.
Schielzeth, Holger & Wolfgang Forstmeier. 2008. Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20(2). 416–420.
Sönning, Lukas & Manfred Krug. 2021. Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In Ole Schützler & Julia Schlüter (eds.), *Data and methods in corpus linguistics: Comparative approaches*. Cambridge: Cambridge University Press.
Sorensen, Tanner & Shravan Vasishth. 2015. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint arXiv:1506.06201*.
Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.
Sóskuthy, Márton & Timo B. Roettger. 2020. When the tune shapes morphology: The origins of vocatives. *Journal of Language Evolution* 5(2). 140–155.
Spina, Stefania. 2019. Role of emoticons as structural markers in Twitter interactions. *Discourse Processes* 56(4). 345–362.
Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. (Textbooks in Language Sciences 7). Berlin: Language Science Press.
Stephan, Johannes, Oliver Stegle & Andreas Beyer. 2015. A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications* 6(1). 1–10.
Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348.
Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150.
Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1). 76–99.
Szmrecsanyi, Benedikt, Biber Douglas, Jesse Egbert & Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28(1). 1.
Tagliamonte, Sali A. & Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
Tagliamonte, Sali A. & Katharina Pabst. 2020. A cool comparison: Adjectives of positive evaluation in Toronto, Canada and York, England. *Journal of English Linguistics* 48(1). 3–30.
Trude, Alison M. & Sarah Brown-Schmidt. 2012. Talker-specific perceptual adaptation during online speech perception. *Language & Cognitive Processes* 27(7–8). 979–1001.
Vul, Edward, Christine Harris, Piotr Winkielman & Harold Pashler. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* 4(3). 274–290.

Wallis, Sean. 2021. *Statistics in corpus linguistics research: A new approach*. New York, NY: Routledge.
Wells, Gary L. & Paul D. Windschitl. 1999. Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin* 25(9). 1115–1125.
Wendorf, Craig A. 2002. Comparisons of structural equation modeling and hierarchical linear modeling approaches to couples' data. *Structural Equation Modeling* 9(1). 126–140.
Westfall, Jacob, David A. Kenny & Charles M. Judd. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* 143(5). 2020.
Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70. 86–116.
Wieling, Martijn, Simonetta Montemagni, John Nerbonne & Harald Baayen. 2014. Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90(3). 669–692.
Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS One* 6(9). e23613.
Winter, Bodo. 2011. Pseudoreplication in phonetic research. In Lee Wai-Sum & Eric Zee (eds.), *Proceedings of the 17th International Congress of Phonetic Science*, 2137–2140. Hong Kong: City University of Hong Kong.
Winter, Bodo. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, (paper number 0181.1-4). Glasgow: The University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0181.pdf.
Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York, NY: Routledge.
Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution* 1(1). 7–18.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.
Yarkoni, Tal. 2020. The generalizability crisis. *Behavioral and Brain Sciences* 1–37. https://doi.org/10.1017/S0140525X20001685. https://psyarxiv.com/jqw35/.