On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally

Stefan Th. Gries
University of California & Justus Liebig University Giessen

This paper critically discusses how corpus linguistics in general, but learner corpus research in particular, has been dealing with all sorts of frequency data in general, but over- and underuse frequencies in particular. I demonstrate on the basis of learner corpus data the pitfalls of using aggregate data and lacking statistical control that much work is unfortunately characterized by. In fact, I will demonstrate that monofactorial methods have very little to offer at all to research on observational data. While this paper is admittedly very didactic and methodological, I think the discussion of the empirical data offered here – a reanalysis of previously published work – shows how misleading many studies potentially and provides far-reaching implications for much of corpus linguistics and learner corpus research. Ideally/maximally, this paper together with Paquot & Plonsky (2017, *Intntl. J. of Learner Corpus Research*) would lead to a complete revision of how learner corpus linguists use quantitative methods and study over-/underuse; minimally, this paper would stimulate a much-needed discussion of currently lacking methodological sophistication.

Keywords: learner corpora, speaker/file variation, multifactorial analysis, over-/underuse

1. Introduction

1.1 General introduction

The sub-discipline of corpus linguistics that studies the use of a language by non-native speakers of that language, learner corpus research, has been steadily growing in size and influence. This field now has its own flagship journal (the *International Journal of Learner Corpus Research*), its own (Cambridge) handbook, its own conference series (LCR, the last conference was in 2017 in Bolzano, Italy),

https://doi.org/10.1075/jsls.00005.gri (proofs)
Journal of Second Language Studies 1:2 (2018), pp. 276–308. issn 2542-3835 | e-issn 2542-3843

and its own association (the Learner Corpus Association, <http://www.learnercorpusassociation.org/>).

These developmental milestones notwithstanding, the field is still young, still coming of age, and still evolving in a variety of ways, just as is the larger field of corpus linguistics of which it is a part. More specifically, learner corpus research is 'in a hard place' because not only does it have all the problems corpus linguistics in general faces, but then also has to come to grips with their own specific challenges. In a recent survey article, Paquot and Plonsky (2017) provide an instructive overview of the many methodological problems corpus linguistics and learner corpus research are still facing. With regard to corpus linguists in general, they observe that corpus linguists too rarely

- inspect results obtained from some corpus for variability that arises from the composition of the corpus, i.e. the fact that the corpus consists of different parts, whose size/complexity ranges from modes over registers and sub-registers down to files that represent one or more specific speakers' contribution to a corpus;
- quantify between-corpus-parts variability corpus homogeneity (see Gries, 2006).

At the same time, corpus linguists too often

- rely on observed frequencies and how they relate to other observed frequencies or expected frequencies, i.e. what in learner corpus research is often referred to as *over- and underuse*; this will play a central role below;
- ignore the role that dispersion can play in corpus-linguistic studies (see Gries, 2008);
- do not consider the role(s) that multiple predictors/causes play with regard to a certain phenomenon, which is often coupled with the fact that corpus-linguistic research also often ignores how multiple predictors behave together – additively or interactively;
- ignore the repeated-measurements and hierarchical structure in their data: subjects routinely provide more than one data point of a certain phenomenon, which rules out the use of many traditionally taught/used statistical methods such as the chi-squared test or the log-likelihood test, and corpora often have a hierarchical structure (files nested into sub-registers nested into registers nested into modes, as in the case of the ICE-GB), which increases the complexity of the required statistical analyses (see Gries, 2015);
- compute multiple statistical (post-hoc) tests on one and the same data set without correcting for that;

– underutilize effect sizes, confidence intervals, proper/insightful visualization, and many other more advanced methods of statistical analysis.

Crucially for the purpose of the present paper, Paquot and Plonsky conclude that many of these criticisms also apply to learner corpus research. More specifically, in a survey of 378 learner corpus studies, about 90% of all statistical analyses are one of the following six, most of which are applications that likely underestimate the complexity of the data learner corpus research is studying: chi-squared tests (23%) and the log-likelihood ratio G^2 (10%), *t*-tests (20%) and simple correlation (17%), analysis of variance (14%) and regression (6%).

From their survey, Paquot and Plonsky derive a variety of improvements that both corpus linguistics in general and learner corpus research in particular would benefit from and many of these have already been integrated into the above discussion: corpus linguists should not aggregate results from different speakers, use the case-by-variable format for data frames, conduct fewer (exploratory/post-hoc) significance tests, control the alpha-level of the significance tests that are conducted, report effect sizes and confidence intervals, consider multifactorial statistics etc.

1.2 The two goals of this paper

To a statistically less experienced reader, the above list of recommendations by Paquot and Plonsky looks nothing less but daunting of not insurmountable – which is understandable. However, there is a way to reframe this because most of Paquot and Plonsky's recommendations really only boil down to do *proper regression modeling* on case-by-variable data (i.e. a matrix-like format in which every observation of the dependent variable gets its own row and is annotated for multiple variables in separate columns). Crucially, most of the statistics in learner corpus research are actually already instances of regression modeling even if most readers will probably not realize that, given the terminology that is used. But actually,

– a *t*-test is a kind of a linear regression model, namely one with a single binary predictor;
– a simple correlation is a kind of linear regression model, namely one with a single numeric predictor;
– an analysis of variance is a linear regression model with one or more binary/categorical predictors;
– a 'regression' in the stricter sense is a linear regression with one or more numeric predictors;

– a log-likelihood G^2-value is the significance test of a binary logistic regression, and a chi-squared test is related to it conceptually and usually highly correlated with it.

In other words, the field of learner corpus research is already doing regressions – it's just not calling it that and, unfortunately, it's not doing them on the right kinds of data and it's not doing them right. Correspondingly, the first goal of this paper is (i) to discuss how and why current statistical practices in particular with regard to over- and underuse do some damage to the field and its results and (ii) to show how this can be done better; this is the topic of some discussion and exemplification in Section 2 and will lead to an account of over- and underuse that essentially meets just about all of Paquot and Plonsky's recommendations.

All that being said, the second goal of this paper is to go one step further and argue that even the more sophisticated way of studying over- and underuse outlined in Section 2 is, while statistically 'legit' and much safer, nevertheless not even an insightful to study learners' over- and underuse. While this view may strike the reader as outrageous, given the dozens of studies based on this very concept, I hope to show that this methodological tradition is nonetheless not a fruitful one; this is the topic of Section 3.

2. A regression-modeling approach to over- and underuse

2.1 The data: *quite* in learner and native-speaker data

In this section, I will discuss how a regression-modeling perspective on over- and underuse is vastly superior to the vast majority of existing work in that area. As an example data set, I will use a part of the data as discussed in Hasselgård and Johansson (2011), who explored the use of the word *quite* in four learner corpora as compared to the use of *quite* in native-speaker data. Specifically, they are comparing *quite*'s frequency in the LOCNESS corpus of native-speaker writing to *quite*'s frequency in two corpora of English learners with a Romance L1 background (Spanish and French) in the ICLE corpus as well as to *quite*'s frequency in two corpora of English learners with a Germanic L1 background (Norwegian and German), also from the ICLE corpus; Table 1 represents their data.

Table 1. Data from Table 2 from Hasselgård & Johansson (2011:46)

	LOCNESS	ICLE-SP	ICLE-FR	ICLE-NO	ICLE-GE
Frequency	67	63	78	92	147
Frequency pmw	205	318	380	437	623

They report that "*quite* is overused in all the learner groups but most markedly so among the Germans, followed at a distance by the Norwegians (both at significance levels of $p < 0.01$)" (p. 45); their footnote 14 then explains that the statistical test used for the frequency comparisons: "The frequencies from each ICLE sub-corpus and LOCNESS were compared using chi-square" (p. 45). Also, they argue that "the overall frequency distribution […] seems to reflect the Germanic – Romance distinction" (p. 45f.).

This part of their paper is highly problematic in how it goes against virtually all methodological recommendations issued in corpus linguistics:

1. it aggregates everything that happens in each corpus into a single observed frequency value, thus ignoring by-speaker/by-text variability;
2. it is what I will call essentially zero-factorial: no causes that may (co-)determine the use of *quite* are explored in this part (their later discussion involves word classes of words after *quite*), but does not integrate that into one compelling statistical analysis;
3. while the exact nature of their statistical testing remains unclear from their note 14 – from their brief description, I was not able to replicate their results with neither chi-squared tests for independence nor chi-squared tests for goodness-of-fit – it does seem clear that they compared each learner variety separately against the native speaker data, i.e. they minimally did four or five tests (EN vs. SP, EN vs. FR, EN vs. NO, and EN vs. GE if not also SP/FR vs. NO/GE) without correcting for this;
4. they provide no effect sizes, no confidence intervals, no visualization.[1]

Several comments are due at this point: First, it is important to note that the point of this paper is *not* me trying to gratuitously trash Hasselgård & Johansson (2011), something that I have assured the first author of in personal communication. I am discussing this paper as detailed as I do because it is a paper that is very representative of a lot of past and current work in learner corpus research; for instance, the following studies all involve similarly aggregated over- and underuse counts and many of them involve similar chi-squared or log-likelihood tests: Hyland & Milton 1997; Aijmer 2002; Altenberg 2002; Connor et al. 2005; Laufer & Waldman 2011; Gilquin & Granger 2011; Neff van Aertselaer & Bunce 2012; and doubtlessly many more; this kind of research is even attested nowadays, as when Gilquin and Lefer (2017) present a study perfectly analogous to Hasselgård and Johansson

1. I am ignoring other, more linguistic, problems such as the fact that the uses of *quite* may contain mistakes such as *in the text there are quite allusions to Pamela* or *their need for peace and quite*, which, to stay as closely to Hasselgård & Johansson's frequencies, I am counting, too.

involving over- and underuse frequencies of negative affixes with learners from two Romance and two Germanic learner varieties.

Second, I am discussing this case study of Hasselgård and Johansson in such detail because the data they use are easily retrievable from the relevant corpora and, as will be shown below, have intriguing characteristics that make for a very instructive discussion.

Finally, I am not discussing these issues just theoretically, because I believe showing the impact of different methods on corpus linguistics is more persuasive when accompanied by real-life examples. It is easy to say, as I have heard many times, that "oh, just another number-crunching talk where he's arguing against a straw man, no one really does that or it doesn't matter", but it is less easy to be that dismissive when a concrete example is provided …

With that background, the next two sections will discuss what happens when we gradually approximates the recommendations of many methodologically savvy corpus linguists as well as Paquot and Plonsky's. Specifically, Section 2.2 discusses the results of a generalized linear model on Hasselgård and Johansson's data, whereas Section 2.4 discusses the results of a generalized linear mixed-effects model.

2.2 A generalized linear model on the Hasselgård and Johansson data

A researcher attempting to implement Paquot and Plonsky's recommendations when replicating Hasselgård and Johansson's study has many options: In this section, I will discuss what happens if one indeed analyzes the data based on the case-by-variable format, limits the number of significance tests, reports effect sizes and confidence intervals, and visualizes. To that end, the following strategy was implemented: First, I wrote an R script that loads every file from LOCNESS (for the native speakers) as well as every file from ICLE-SP, ICLE-FR, ICLE-NO, and ICLE-GE. Each of these files was then split up into words (using a simple heuristic regular expression, namely "[^'A-Za-z]+", i.e. any sequence of characters not a small or capital letter between *a* and *z* or an apostrophe). Then, for each file, a vector was created that was as long as the file has words and that consisted of *yes* for every word that was *quite* and *no* for every word that was not. That means, the file ICLE:GEAU2039 was represented by a vector consisting of 171 *no*'s and 1 *yes*'s, because it has 172 words as defined above and one of them is *quite*, in the sentence *But what happens quite often?*. Then, in the final data preparation step, I created one data frame called x containing all the corpus data such that

- column 1 stated for every word in all of LOCNESS and the four ICLE parts whether it was *quite* or not;
- column 2 stated for every word in all of LOCNESS and the four ICLE parts which file it was from;
- column 3 stated for every word in all of LOCNESS and the four ICLE parts what the L1 of the author of the file was.

This is represented here in R notation: first a summary of the data frame x, then its first rows, then its last rows:[2]

```
> summary(x)
 QUITE                FILE           L1
 no :1197086    BRSUR1_01:   3704    EN:323898
 yes:    451    BRSUR1_05:   3545    SP:199789
                SPAL1005 :   3395    FR:227361
                BRSUR1_15:   3161    NO:213716
                BRSUR1_04:   3071    GE:232773
                BRSUR1_09:   2995
                (Other)  :1177666

> head(x, 3)
 QUITE   FILE L1
1 no FRUB1001 FR
2 no FRUB1001 FR
3 no FRUB1001 FR

> tail(x, 3)
        QUITE     FILE L1
1197535   no USMIXED33 EN
1197536   no USMIXED33 EN
1197537   no USMIXED33 EN
```

This way, we are following Paquot and Plonsky's (correct) recommendation to use the case-by-variable format and are at the same time following the Principle of Accountability (Labov 1982:30): each lexical choice, *quite* or not *quite*, enters into the analysis. And, as the following two lines show, we are reasonably – in fact, *very* – close to the results Hasselgård and Johansson report both in terms of absolute and relative frequencies; the slight discrepancies are immaterial to the methodological points of this paper:

```
> with(x, table(QUITE, L1))[2,]
 EN   SP   FR   NO   GE
 67   64   82   92   146
> round(prop.table(with(x, table(QUITE, L1)), 2)[2,]*1E5, 1)
  EN    SP    FR    NO    GE
20.7  32.0  36.1  43.0  62.7
```

2. R output may be abridged to omit details irrelevant to the current discussion.

This format now means we could use a generalized linear model, a binary logistic regression, to be exact, to model the occurrences of *quite* in all corpora as a function of the L1 of the speaker, which could be written as follows:

```
> summary(model <- glm(QUITE ~ L1, data=x, family=binomial))
```

This would return a summary table that provides

- an overall significance test in the form of a G^2-value ($G^2 = 65.475$, $df = 4$, $p < 10^{12}$);
- four coefficients that contrast the occurrence of *quite* in the native-speaker data with the occurrence of *quite* in each learner variety: all of these coefficients are positive (indicating that the learners have higher occurrences of *quite* than the native speakers) and all of these coefficients are significant (all $p < 0.013$, indicating that these differences are not that compatible with the null hypothesis of no differences between native and non-native frequencies of *quite*).

Note how this is already quite an advantage in how we arrive at one omnibus significance test and the coefficients of the regression model embodying that one omnibus test provide one kind of perspective one might be interested in, comparing each learner variety to the native speaker data. In addition, and that is already something very very few learner corpus studies provide, we also obtain (i) an assessment of how well the regression model can distinguish between uses of *quite* and other words and (ii) the results from such a model allow us to compute effect sizes (e.g., odds ratios) and confidence intervals, which in turn allows us to visualize the results well. It will be important for later to already point out how low the classification power of the regression model is: Nagelkerke's R^2 is as small as 0.008 and the *C*-value of this model is at its theoretical minimum of 0.5. The odds ratios for the comparisons of the four learner varieties and the native-speaker data are 1.55 (for SP, 95%-*CI*: (1.1, 2.18)), 1.74 (for FR, 95%-*CI*: (1.26, 2.41)), 2.08 (for NO, 95%-*CI*: (1.52, 2.86)), and 3.03 (for GE, 95%-*CI*: (2.28, 4.07)), and Figure 1 represents the predicted probabilities of *quite* per corpus.

However, there is one other tweak we can make to make the model address even more specifically some of the conclusions that Hasselgård and Johansson made. Regression models allow users to not just obtain the standard (treatment) contrasts that just about all corpus-linguistic studies are reporting (and that are in fact correlated with each other), but also to define and test user-defined orthogonal contrasts, i.e. coefficients that test/answer user-defined questions and are even nicely independent of each other (and offer other advantages as well regarding multicollinearity and the interpretation of main effects). In this case it so happens that the four orthogonal contrasts one can define for a predictor with five levels

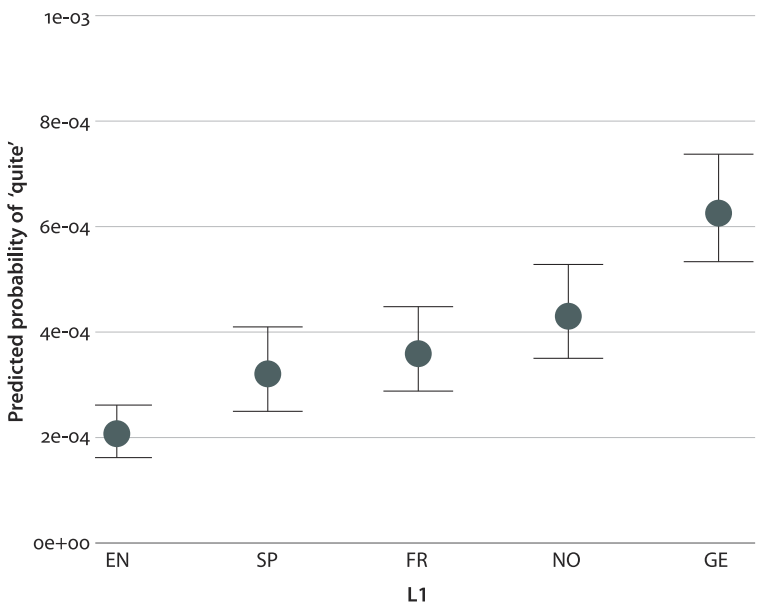

Figure 1. Predicted probabilities of *quite* resulting from a glm on LOCNESS and the four ICLE parts

such as L1 are all of interest; the most useful contrast coding would be this, which is then also illustrated in Figure 2.

- Contrast 1 (dark grey): EN vs. (SP & FR & NO & GE), i.e. is there a difference between the native vs. all non-native data combined?
- Contrast 2 (light blue): (SP & FR) vs. (NO & GE), i.e. is there a difference between the Romance vs. Germanic L1s?
- Contrast 3 (orange): SP vs. FR, i.e. is there a difference between the Romance L1s?
- Contrast 4 (pink): NO vs. GE, i.e. is there a difference between the Germanic L1s?

All general statistics (G^2, R^2, …) stay the same, but the coefficients change:

```
> summary(model.l1s.glm <- glm(QUITE ~ L1, family=binomial, data=x))
                    Estimate Std. Error z value   Pr(>|z|)
(Intercept)          -7.91602    0.04919 -160.935  < 2e-16 ***
L1native_vs_nonnative -0.70909    0.13333   -5.318 1.05e-07 ***
L1rom_vs_germ         -0.42463    0.10672   -3.979 6.92e-05 ***
L1span_vs_fren        -0.11860    0.16682   -0.711  0.47713
L1norw_vs_germ        -0.37660    0.13315   -2.828  0.00468 **
```

Leaving aside the intercept, which is not relevant here, the remaining four coefficients address the four contrasts defined above, indicating that (i) the fre-

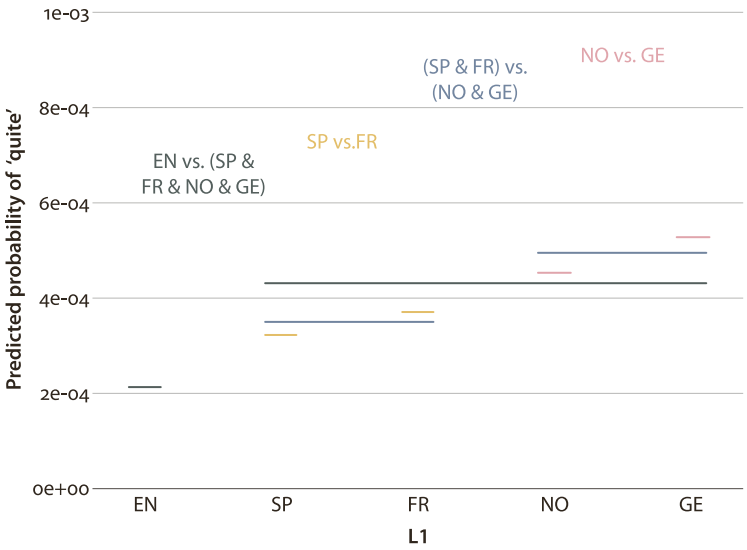

Figure 2. Visualization of orthogonal contrasts

quencies of *quite* in the native and the non-native data (as a whole) are significantly different from each other; (ii) the frequencies of *quite* in the Romance and the Germanic data (each as a whole) are significantly different from each other (which appears to support Hasselgård and Johansson); (iii) the frequencies of *quite* in the Spanish and the French data are not significantly different from each other (which appears to support Hasselgård and Johansson's idea that the Romance data are homogeneous enough to compare them to the Germanic data); (iv) the frequencies of *quite* in the Norwegian and the German data are significantly different from each other (which appears to *contradict* Hasselgård and Johansson's idea that the Germanic data are homogeneous enough to compare them to the Romance data).

We have made quite some progress with regard to Paquot and Plonsky's recommendations: to recap, we are now using the case-by-variable format, we are doing only planned and required significance test that directly and without exception test hypotheses of interest, we are computing and visualizing effect sizes as well as confidence intervals; also, we are assessing the overall quality of the statistical model (with R^2 and C), something which most over- and underuse studies have not done. However, while this is quite some improvement and already introduced quite some necessary complexity that many previous over- and underuse studies did not involve, this is still not sufficient for reasons to be discussed in the following section.

2.3 Why a generalized linear model is not enough

The main problem of the previous analysis, and by extension of any analysis that is even simpler, is that it does not take the distribution of the data seriously enough into account because speaker-specific variability is not considered at all. Figure 3 visualizes why that is a problem: The L1s of the speakers are on the *x*-axis, relative frequencies of *quite* are on the *y*-axis, the colored lines are those of Figure 2, and, most importantly, each file's percentage is represented separately by a jittered olive-green point; in addition, the values printed at $y = 0.83$ represent the relative frequencies of corpus files per L1 that do not even contain *quite*.

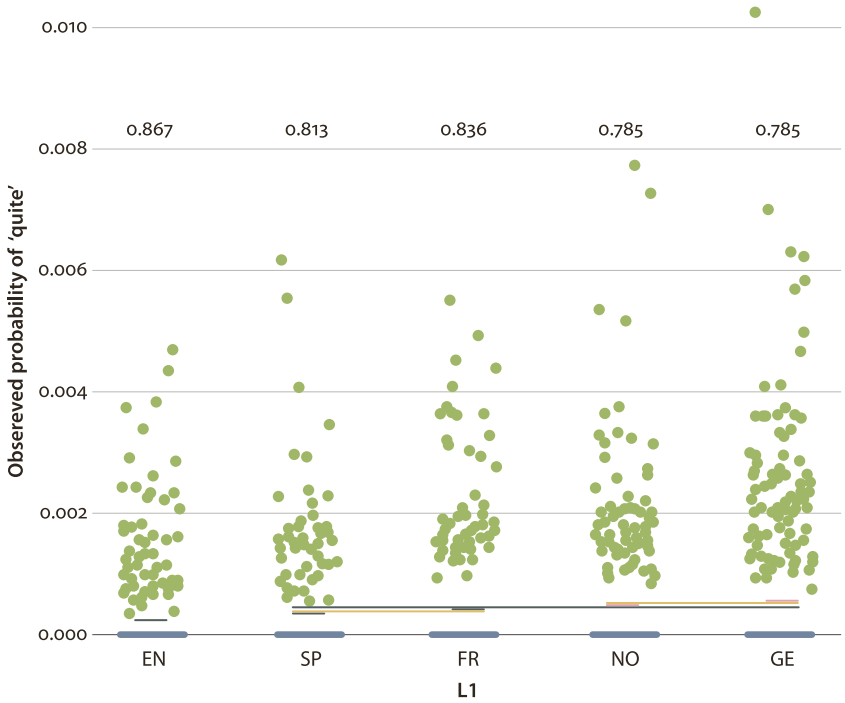

Figure 3. The frequencies of *quite* in LOCNESS and the four ICLE components in question on a by-file/speaker basis

Several findings emerge from this: First, the actually observed percentages of *quite* are very small and very much located at the bottom of this plot. Second, this is due to the facts that (i) for every L1, there is a large amount of variability as indicated by the many olive-green points above the per-L1 means and (ii) for every L1, the vast majority (on average, >80%) of all files does not contain *quite* even once. The huge variability and the fact that most speakers do not use even

use *quite* are characteristics of the data that any analysis based on aggregate data simply ignores.

2.3.1 *Which files to include?*

This raises a question, that of whether or not to include files in which the linguistic unit in question is not even attested. This question is one that is, to my knowledge at least, not discussed or appreciated enough, but is both important and tricky: If, as is usually the case, files without the unit in question are included in the analysis, then this potentially introduces some vagueness into the analysis because the absence of the unit in question can be an intentional choice to not use it, or it can result from the fact that the learner does not know the unit at all and so does not know there is a choice to be made. For instance, in a study of the dative alternation between *John gave Mary the book* and *John gave the book to Mary* a learner might not pick the ditransitive because everything in the relevant context screams out for a *to*-dative or because he does not even know a ditransitive exists.

In the former case, one would obviously want to include the speaker's data in the analysis because the speaker's choices provide important information regarding which construction is chosen under what circumstances. Obviously, in the latter case the situation is different and it is less obvious what to do because a speaker who does not know there is a ditransitive and will always use the prepositional dative, even if everything in the situation – say, the verb *give* used with a human agent, a given short human recipient, and a long concrete patient – requires a ditransitive, will provide input to the analysis that not only skews things statistically, but also throws off any conceptual interpretation because of how it downplays the effect of factors that promote the unit he does not know. It seems to me as if analysts always include all speakers and just assume that the unit(s) in question will be known to the speaker and that not using (one of) it is a choice that does not just derive from not knowing any alternatives. The probability of this assumption being correct is extremely high for the native speakers that may figure in a comparison to learners, and the probability of it being correct is positively correlated with the the L1 background and the proficiency of the learners. For instance, German learners with a B2, C1, or C2 CEFR level of English can be assumed to know about a ditransitive construction, whereas native speakers of a language without ditransitives who are A2 learners of English maybe cannot. It may be possible to develop such assumptions from existing corpora by (i) determining which units learner *A* knows and doesn't know, (ii) checking what other learners with a comparable L1 background and proficiency know and do not know, and then (iii) impute whether learner *A* is likely to know unit *x*: if many learners comparable to *A* know *x*, we can compute how likely it is that *A* knows it, too.

While I have no clear-cut solution to offer to this problem, it is one that I think requires much more discussion and operationalization. This is because including these choices have some implications. The following regression output shows the result when the regression model is restricted to only the files that contain at least one *quite*, …

```
> summary(model.l1s.glm.b <- glm(QUITE ~ L1, family=binomial))
                     Estimate Std. Error   z value  Pr(>|z|)
(Intercept)          -6.34698   0.04922   -128.950   < 2e-16 ***
L1native_vs_nonnative -0.51027   0.13339     -3.825  0.000131 ***
L1rom_vs_germ         -0.19649   0.10680     -1.840  0.065798 .
L1span_vs_fren        -0.42771   0.16694     -2.562  0.010406 *
L1norw_vs_germ        -0.13490   0.13325     -1.012  0.311358
```

and the next output shows the result when the regression is based on all native-speaker files and all learner files containing *quite* at least once:

```
> summary(model.l1s.glm.c <- glm(QUITE ~ L1, family=binomial))
                     Estimate Std. Error   z value  Pr(>|z|)
(Intercept)          -6.69260   0.04922   -135.986   <2e-16 ***
L1native_vs_nonnative -2.23836   0.13334    -16.787   <2e-16 ***
[…]
```

The difference between native and non-native speakers remains significant, but that is unsurprising given the huge sample sizes involved here, but the effect sizes change massively. Also, in both analyses (i) the significant difference between the Romance and the Germanic learners disappears, (ii) the formerly insignificant difference between Spanish and French learners now becomes significant, and (iii) the formerly significant difference between Norwegian and German learners becomes insignificant. In other words, most results are massively different depending on which of the learner files are included – only those with *quite* or all of them.

2.3.2 *The role of individual speakers*

There is a third finding emerging from Figure 3 that merits discussion, namely the one German speaker with the highest percentage of *quite* (> 0.01) in the top right corner. In the interest of brevity, I will not discuss this issue in detail (with regression coefficients or significance tests), it is worth pointing out that removing just this single German learner from the data alters the results for all the totality of the German learners using *quite* considerably such that, for instance, suddenly the German learners are not the ones with the highest frequencies of *quite* anymore.

To nevertheless showcase the danger of not taking file-/speaker-specific distribution, i.e. dispersion, seriously, consider the following situation. Imagine you are researching the genitive alternation between *of*- and *s*-genitives (as in [$_{\text{possessed}}$ *the nuts*] *of* [$_{\text{possessor}}$ *the squirrel*] vs. [$_{\text{possessor}}$ *the squirrel*]'s [$_{\text{possessed}}$ *nuts*]) and you are suspecting (correctly) that, because of the well-known short-before-long

preference that characterizes many English alternations, the length difference between possessor and possessed helps predict which genitive choice a speaker will make. To determine the nature of this correlation, you look at a corpus based on 20 speakers and plot the genitive choices against the difference $\text{length}_{\underline{\text{possessor}}}$ minus $\text{length}_{\underline{\text{possessed}}}$; this might result in Figure 4, which represents a highly significant correlation with a Spearman's *rho*-value of 0.5.

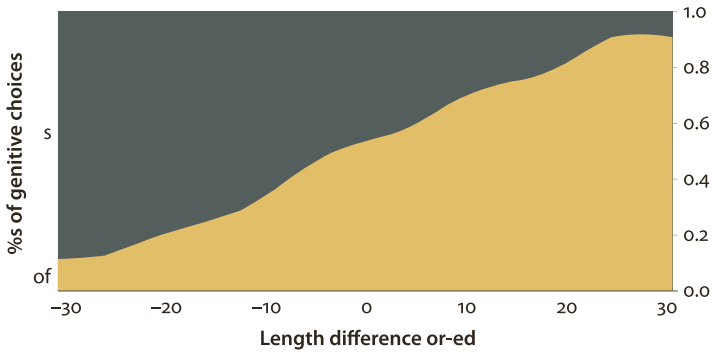

Figure 4. A conditional density plot of genitive choices against or-ed length differences

As everyone who is familiar with Simpson's paradox knows, the scary thing about this data is that this strong overall correlation can arise from very different speaker behaviors. On the one hand, this overall correlation of *rho* = 0.5 can arise from 20 speakers most of whom behave like the summary in Figure 4. That situation is depicted in Figure 5: the *x*-axis represents the 20 speakers, the *y*-axis represents the Spearman rank correlation between genitive choices and or-ed length differences for each speaker, and the horizontal line with the grey bar represents the mean rank correlation and its 95%-confidence interval, which precisely includes the correlation value of the aggregated data in Figure 4.

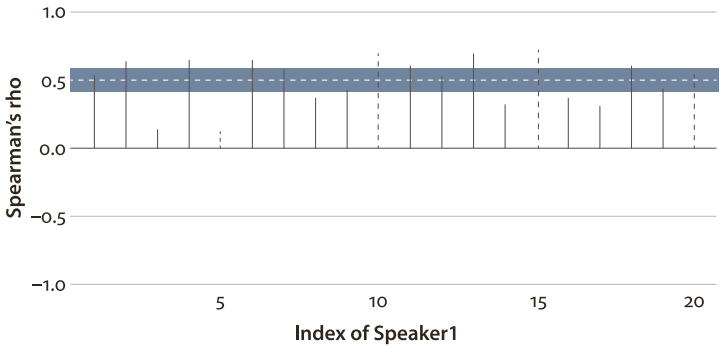

Figure 5. Spearman rank correlations of 20 speakers giving rise to the data in Figure 4

On the other hand, that Figure 4 can also arise from 20 speakers hardly any of whom behave anything like the overall correlation suggests: nearly all speakers exhibit correlations between genitive choices and length differences with an absolute value of <0.3 and an overall average Spearman rank correlation of −0.03 with a confidence interval including 0.

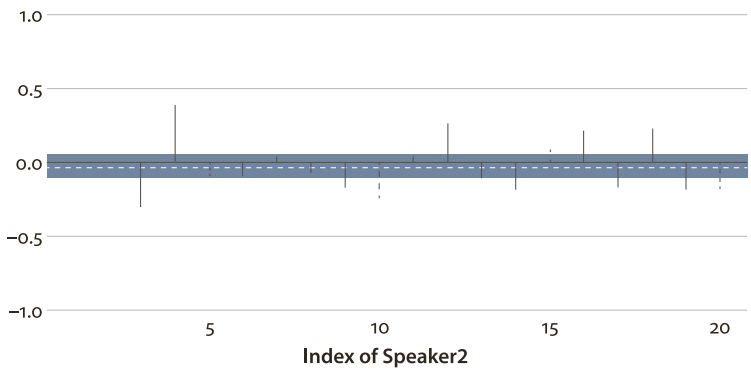

Figure 6. Spearman rank correlations of 20 different speakers also giving rise to the data in Figure 4

Again, this danger is no news to anyone who is familiar with Simpson's paradox: results from an aggregated data set can disappear or even be reversed when the same data set is split up into subgroups. However, in spite of the widespread presence of Simpson's paradox in statistical textbooks (see Gries, 2013, pp. 5–6, Section 5.3.5 for two brief examples), it is probably fair to say that the majority of learner corpus studies has not protected itself properly against the possibility that their analyses completely underestimate the variability in the data and the leverage that very small numbers of speakers can have on their data. Therefore, the final part of Section 2 proceeds with a discussion of how regression modeling of the above type can be extended to address – at least to some extent – the issue of speaker-specific variation.

2.4 A generalized linear mixed-effects model on the Hasselgård and Johansson data

The answer to at least some of the issues is to move from generalized linear modeling to generalized linear mixed-effects modeling, which addresses the violation of the general linear model's assumption that the data points are all independent, which they are not since many speakers provide more than one data point. Admittedly, in this particular data set, the advantages of this approach will be

more limited/theoretical because *quite* is not exactly a frequent word and most speakers do not use it all. However, with many other phenomena, especially of course more frequent words or grammatical patterns/constructions, the advantages will be much more sizable and the added complexity in coding at least is so low anyway that there is no good reason not to use the better approach. Fitting the simplest kind of this model on our current data leads to the following output for the fixed effects (with our orthogonal contrasts).

```
> summary(model.l1s.glmer <- glmer(QUITE ~ L1 + (1|FILE),
 family=binomial, data=x))
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.7274   0.1137   -76.79  < 2e-16 ***
L1native_vs_nonnative -0.6513   0.1667   -3.91  9.36e-05 ***
L1rom_vs_germ         -0.3940   0.1442   -2.73    0.0063 **
L1span_vs_fren        -0.1004   0.2219   -0.45    0.6508
L1norw_vs_germ        -0.2807   0.1849   -1.52    0.1289
```

This model again models the occurrence of *quite* based on the L1 of the speakers, but the random-effects term (1|FILE) states that each speaker can have his own intercept, i.e. baseline of using *quite*, whereas the previous models were less flexible and 'forced' one intercept/baseline on all the speakers. This result is interesting for several reasons. First, because it actually completely supports Hasselgård & Johansson's original results: There is a difference between the native and the non-native speakers and there is a difference between the Romance and the Germanic L1 speakers, but there are no significant differences between the Spanish and the French speakers or between the Norwegian and the German speakers. However, and I cannot emphasize this enough, this does *not* mean that their analysis was sufficient – it means they were lucky that a more appropriate (and complex) analysis happens to support theirs. This is because there is no way any researcher can look at aggregate corpus frequencies or even something as detailed as Figure 3 and determine by eyeballing what the real results are like, and that is especially true in cases where the number of uses of the unit in question per speaker are much higher than here so that anticonservative results are much more likely.[3] More specifically, in this data set, the changes resulting from the mixed-effects model are so small because only 1.5% of the files contain more than one *quite* – but (i) Hasselgård and Johansson did not know that (at least they do not report that), (ii) other linguistic units *will* be much more frequent per speaker so the results would change much more, and (iii) even in this case, note that, once the right kind of analysis is done, all of the user-defined contrasts move in the direction of 0, meaning their effects become weaker, meaning the previous analysis overestimated the magnitude of the effects. This is especially true if the analysis is restricted to only

3. Results are anticonservative if they increase type-1 errors, i.e., make accepting the alternative hypothesis (much) more likely than the nominal significance level (of usually 5%) would permit.

those files that contain *quite* at least once: then, the effect of the contrast between Romance and Germanic L1s shrinks by 50% and the effect of the contrast between Spanish and French increases by a factor of four.

Second, this result is interesting because even with the statistically most sophisticated replacement of the traditional over- and underuse methods, the equivalent of the Nagelkerke R^2-value is ridiculously small, 0.027. This leads to the first of two conclusions that should be very disconcerting to anyone who wants learner corpus research to succeed: All learner corpus research studies based on over- and underuse frequencies that result from aggregating corpus data and that are tested with log-likelihood tests (or, worse, chi-squared tests) need to be redone because (i) we don't know whether these authors will be as fortunate as Hasselgård and Johansson to have their results confirmed and (ii) we do know that aggregation and multiple testing of the types that have been rampant skew results in an anticonservative direction. Unfortunately, this first conclusion will be qualified further below.

3. Multifactoriality: What it means for over- and underuse and in general

From the last section, it would seem as if over- and underuse studies 'are now safe': ok, so corpus linguistics and learner corpus research must not pursue the traditional kind of studies using aggregate corpus data and chi-squared/log-likelihood tests anymore, but with generalized linear mixed-effects models we can now address everything that is problematic about these old-style tests.

Unfortunately, this is not the whole truth. Yes, the mixed-effects modeling approach addresses nearly all frequent statistical points of critique but the first hint at a problem remaining is the fact we still have only very small R^2-values. But, and this is the final twist in this paper, there is an additional complication which is not methodological, but epistemological and linguistic/conceptual in nature and we will need to make a slight detour to cover it.

3.1 Why virtually every corpus study, every one, needs to be multifactorial

The fact of the matter is this: We are trying to understand a linguistic phenomenon, here the use of *quite*, but it could be anything else such as the genitive alternation mentioned earlier. The use of *quite*, the use of *of*- vs. *s*-genitives, or the use of any other linguistic unit is probabilistically shaped by probably very many different factors, i.e. is a potentially multifactorial phenomenon. The complication now is that over- or underuse studies attempt to tackle a phenomenon that is multifactorial on the basis of a monofactorial test – even the mixed-effects model dis-

cussed in the previous section involves only *one* predictor, L1 – but, to put it as boldly as well as as clearly as possible: monofactorial observational studies have virtually *nothing* to contribute to corpus linguistics!

Let me explain why I dare make such a bold claim, essentially laying waste to much corpus linguistic work, using an easily comprehensible non-linguistic example first, because this kind of example will arouse little theoretical disagreement. This example is concerned with the efficiency of cars as measured by their mpg (miles per gallon) value, i.e. how many miles (1.609 km) can a car travel on one gallon (3.79 liters) of gas. Imagine you think and read about this topic and you find a study from the early 1990s that shows that cars with more cylinders need more gas (i.e. have a lower mpg-value). Imagine there's also another study from the late 1990s that shows that cars with more horsepower have lower mpgs, plus it follow from basic physics that heavier cars would have lower mpgs. But then you get an idea, namely the alternative hypothesis that cars with more displacement should have lower mpgs. That seems reasonable and you do a statistical test of your hypothesis on a range of 32 different cars:

```
> summary(test.of.new.hyp <- lm(mpg ~ disp, data=mtcars))
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 29.599855   1.229720   24.070  < 2e-16 ***
disp        -0.041215   0.004712   -8.747 9.38e-10 ***
Multiple R-squared: 0.7183,  Adjusted R-squared: 0.709
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

You had a monofactorial hypothesis, you did a monofactorial test of it, it turns out highly significant and explains quite a large amount of the variability of the mpg values in your data (>70%) – it seems like it's time to write this up and await a congratulatory acceptance letter.

Not so … This is because mpg values are a multifactorial phenomenon and your test was a test of your alternative hypothesis that disp is correlated with mpg *against the null hypothesis that it is not*. However, that way your test was completely anticonservatively stacked in your favor because you did *not* test your hypothesis *against everything else we already know* to play a role. Put differently, you tested your hypothesis pretending we have no prior knowledge about mpgs (leaving all of the variability of the mpg values, their variance, up for grabs by the one predictor you want to show is important: disp). But in fact we do have some prior knowledge: we *know* that the number of cylinders, horsepower, and weight play a role, which statistically means they already account for a lot variability/variance. In fact, everything we already know about mpg values accounts for >85% of the mpg variability:

```
> summary(prior.knowl <- lm(mpg ~ (cyl+hp+wt)^2, data=mtcars))
             Estimate Std. Error t value Pr(>|t|)
[…]
```

```
Multiple R-squared:  0.895,   Adjusted R-squared:  0.8697
F-statistic:  35.5 on 6 and 25 DF, p-value: 4.665e-11
```

That means what you would really need to test is not whether disp does *anything* (as opposed to nothing) but whether disp *adds to or replaces what we already know*, which means we need to determine (i) the impact that disp has on the variability that we do not already account for with other things and (ii) whether that impact is significantly different from 0. As it turns out, disp has nothing to contribute on top of what we already know: Adding disp and all its pairwise interactions with other predictors to the previous model makes no useful contribution: adjusted R^2 in fact goes down, not up, and disp and all its interactions do not add to the model significantly ($p > 0.99$ and the prior knowledge model is >7500 times as likely to be the 'right model' than the one that also adds disp, as indicated by the evidence ratio):

```
> summary(real.test.of.new.hyp <- lm(mpg ~ (cyl+hp+wt+disp)^2, data=mtcars))
            Estimate Std. Error t value Pr(>|t|)
[…]
Multiple R-squared:  0.896,   Adjusted R-squared:  0.8464
F-statistic:  18.08 on 10 and 21 DF, p-value: 3.773e-08
> anova(prior.knowl, real.test.of.new.hyp, test="F")
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    25   118.28
2    21   117.16 4    1.1232 0.0503 0.9949
> exp((MuMIn::AICc(real.test.of.new.hyp) - MuMIn::AICc(prior.knowl))/2)
[1] 7535.831
```

Why is that? It is because your new predictor disp is so highly correlated with all previous ones ($R^2 > 0.9$) that whatever disp does is already accounted for by our prior knowledge – basically your idea that disp is important was just an idea about how to operationalize differently what we already know:

```
> summary(what.accounts.4.disp <- lm(disp ~ (cyl+hp+wt)^2, data=mtcars))
Multiple R-squared:  0.9257,  Adjusted R-squared:  0.9079
F-statistic:  51.91 on 6 and 25 DF, p-value: 6.558e-13
```

Alternatively, you might check whether disp might not *add to* what we already know, but it might *replace* what we already know – but it turns out it cannot, because our prior knowledge is >3200 times as likely to be the 'right model' than the one consisting only of disp:

```
> exp((MuMIn::AICc(test.of.new.hyp)  - MuMIn::AICc(prior.knowl))/2)
[1] 3227.457
```

In other words, the operationalization idea you had regarding displacement is also not better than the previous ones. Essentially, adding displacement here is the equivalent of explaining particle placement (*John picked up the book* vs. *John picked the book up*) with regard to the length of the direct object in morphemes and then having the 'new idea' that the length of the direct object in syllables

might be an interesting new predictor … Now you might say, 'ok, but these are extreme and unrealistic examples, no one in linguistics would do that', but that's not true. There are many studies on alternations such as the genitive alternation, the dative alternation, particle placement in native and non-native language. Many of these studies proposed that predictors such as the length of the relevant constituents play a role for such ordering choices (e.g., short-before-long); others argued in favor of discourse-functional factors such as the givenness of the referents of the relevant constituents (e.g. given-before-new); yet others suggested that factors such as definiteness of NPs plays a role (definite-before-indefinite) … However, clearly these kinds of predictors are all strongly interrelated: given referents tend to be encoded with short definite or pronominal NPs and new referents tend to be encoded with longer, maybe indefinite and lexical NPs. Similarly, in his 1994 book, Hawkins argued against the importance of discourse-functional determinants of alternations of the above kind suggesting instead that weight-based factors (which are highly correlated with constituent lengths) are have a higher degree of predictive power. However, while a comparison of the absolute correlation strengths of discourse-functional and weight-based factors may or may not show one of the two to be stronger, the fact of the matter still is that the two will be related along the lines discussed above, which means the two explanations are not on the same level: arguably, discourse-functional factors and weight-based factors are not competing for co-determining an alternation because the former are partly responsible for the latter. In sum, monofactorial studies of observational data have nothing to contribute to corpus linguistics because (i) no phenomenon is monofactorial and (ii) even if one had a new *mono*factorial hypothesis of a phenomenon, it would still require *multi*factorial testing to determine either (a) whether it either adds anything to what we already know about the phenomenon (by statistically controlling for what we already know) or (b) whether it replaces (parts of) what we already know about the phenomenon.

How does this inform our discussion of over- and underuse of *quite*? It does so in two ways: First, we need to face the fact that the use of *quite* is a multifactorial issue that a monofactorial test – regardless whether it's an overly simplistic chi-squared test or a generalized linear mixed-effects model – cannot possibly do justice to it, and that is the reason for the pitiful R^2-values of such accounts (as mentioned above at the end of Section 2.4).

Second, simple over- and underuse models with L1 as the only predictor fare badly in both respects: they do not account for much variability on their own and they do not necessarily add much explained variability once other general and reasonable all-purpose predictors are already included. For instance, it makes sense to assume that the chance of *quite* being used, or nearly anything else being

used, is higher if an essay is just longer; also, and to use quick and dirty fixes for proficiency, it makes sense to assume that *quite* is used more if the essay is more lexically diverse and if the words in it are less evenly dispersed in a general native-speaker corpus. If we test this hypothesis by computing for every file separately (so as to deal with the repeated-measurements issue) the relative frequency of *quite*, the essay length, and the lexical diversity of the essay (Yule's *I*), and then compute linear models (with the by-file statistics as units of analysis) and the relative frequency of *quite* as the dependent variable, then we obtain the results represented in Table 2.

Table 2. Comparison of L1 to other common-sense/general-purpose predictors

Data	Predictors	% variance accounted for
all files	L1	0.01383
	essay length + lex. diversity	0.01276
	essay length + lex. diversity + dispersion	0.0227
all files w/	L1	0.06148
1+ *quite*s	essay length + lex. diversity	0.2294
	essay length + lex. diversity + dispersion	0.2514

In other words, the kind of general-purpose predictors that one would actually always want to include – is there any good reason one would not want to control for essay length or some form of proficiency ever, especially when that means one treats files/speaker as a unit of analysis and thereby address the repeated measurements per speaker? – perform pretty much just as well or even much much better than L1. But what if we let a model selection process decide whether including L1 makes a regression model account for uses of *quite* better? An *AICc*-based bidirectional model selection process on all files leads to a model that contains L1 as a predictor, but with the tiniest of effect sizes (partial eta-squared = 0.009); an analogous process on all files with at least one *quite* leads to a model that does not even include L1 (but essay length (as a polynomial to the 2nd degree) and dispersion.

In other words, even controlling for just the barest minimum of file/text-specific information leaves pretty much nothing for L1 left to explain, and we have not even touched many of the individual-variation kinds of variables that are often used in sophisticated SLA research, such as those capturing aspects of personality, aptitude, motivation, and others. Thus, over- and underuse studies of the still most frequent type in the literature may have revealed significant effects when done as tested, but, as we have seen above, their explanatory power (R^2) is so low that in any other kind of study they would not even be considered worthy of writing up,

and without having tested other predictors at the same time it is impossible to say whether they would in fact survive closer scrutiny at all … That in turn leads to a the second equally disconcerting conclusion: Counter to what I said above, it's not even enough to redo all previous over- and underuse studies with generalized linear mixed effects models as exemplified above: While that addresses statistical concerns, it would still lead to statistically correct, but hopelessly monofactorial, studies of multifactorial phenomena that do not even control for the very basic general factors we know are at work. To state it blatantly clearly, over- and underuse studies that do not control for even the most basic things such as length, proficiency/lexical diversity, or dispersion can be their very design not be certain that whatever variability/variance they ascribe to L1 is not in fact the function of something more general. We need something more precise, an approach that addresses many factors at the same time, ideally all the predictors that we know are correlated with a phenomenon, and such an approach will be outlined and (imperfectly and summarily) exemplified in the following section.

3.2 The role of other predictors

Two strategies seem promising when it comes to addressing the complexity that learner corpus research is facing all the time. Both of these involve regression modeling – one is of a traditional kind (not meant in any negative sense), one is a more recent approach that has been developed specifically for learner corpus research but is now also being used in research on indigenized English varieties (and could be used in several other fields).

3.2.1 *Traditional multifactorial regression modeling*

The first and more usual approach is essentially just an extension of what we saw above, namely a multifactorial regression model; see Gries & Deshors (2014, Section 3) for earlier and more detailed discussion. *Ideally*, that model would have the following characteristics (*ideally* also implying that I at least do not know of a single study doing it all and – full disclosure – that includes any and all of my own previous work):

- it is run on data in a case-by-variable format where instances of use of a unit or one of a set of units are annotated for many different variables (see below);
- its dependent variable is most likely binary or categorical, such as *quite* vs. not *quite* for a case of lexical over- or underuse, or *of*- vs. *s*-genitive for a grammatical alternation;
- it features a variety of contextual linguistic predictors of the lexical or the grammatical choice; for many grammatical alternations that would involve

morphological, syntactic, semantic, information-structural/discoursal parameters of the context in which the lexical/grammatical choice is made; again ideally, this would also always include priming effects (which means random sampling from a corpus would have to be done for complete files/speakers, not randomly from the whole corpus);
- it features speaker-specific characteristics such as personality, aptitude, motivation, proficiency (ideally longitudinally), and, crucially, his L1;
- it features text-specific characteristics such as lexical, grammatical, etc. complexity, register/genre information as well as information, where applicable, on the hierarchical structure that a corpus may come in;
- relevant expected differences of interest would be encoded in orthogonal contrasts and the analysis of some 'final model' would involve effect sizes, confidence intervals, significance tests, and visualization.

On the fixed-effects side of things, the regression model would then be defined as involving, minimally!, interactions of all linguistic predictors with L1 and proficiency scores (to determine which predictors differ across different L1 backgrounds and different degrees of proficiency). On the random-effects side of things, it would involve, minimally!, a random structure that describes to the model the (repeated-measurements but also other) structure of the corpus, which would minimally involve varying intercepts for speakers (to capture speakers' overall baselines of use, which can help compensate for speakers not using a word at all or using only one construction from a set of alternatives), but could also include a more complex structure that describes which files are part of which corpus parts etc.

Such a model – a first model testing only relevant hypotheses, a final model of a selection process, an amalgamated model from multimodel inferencing (see Burnham & Anderson, 2002), … – would then *ideally* be evaluated overall using some version of an R^2-value, a classification accuracy, a prediction accuracy based on cross-validation (e.g., 10-fold cross-validation), a *C*-score as well as precision/recall scores. Fixed-effects predictors would be evaluated using *AICc*-scores as well as with effect sizes, confidence intervals, and effects plots, where by effects plots I mean visualizations such as those in Figure 1 above, namely plots of predicted probabilities/values of the model (rather than observed percentages) (see Fox, 2003): this is because only the former, *but not the latter*, control for the effect of everything else in a model, which is why plots of observed percentages are often useless or even misleading.

Finally, such a model would involve at least a brief check on, or exploration of, any random-effects structure as a sanity check, for model validation, and to find potentially interesting patterns in the random effects for future analysis.

This is one kind of analysis that would be able to really address over- and underuse because it would respect repeated measurements and the structure of the corpus, because it would control simultaneously for many things we know affect a certain lexical or grammatical choice, because it would quantify the role everything plays, because it would flag differences between different L1 backgrounds as interactions of any predictor(s) with L1 while already having controlled for effects that are merely due to proficiency, register, or speaker/lexical idiosyncrasies.

3.2.2 *Multifactorial Prediction and Deviation Analysis Using Regressions (or other classifiers)*

The second and more recent approach is one first developed in Gries & Deshors (2014) and Gries & Adelman (2014). This approach tries to make the analysis a bit more precise but also more focused. The above regression approach might return certain effects as significant even if they do not lead to actually different choices. For instance, in one condition, native speakers might be predicted to choose an *of*-genitive with a predicted probability of 70%, whereas certain learners might be predicted to choose an *of*-genitive with a predicted probability of 76%. Now, a *t*-score in a regression model might flag this as a significant difference, but note that both the native and the non-native speakers *are* predicted to use an *of*-genitive. That is, the speakers differ in the strength of the prediction, which scholars interested in probabilistic grammar have argued to be interested in, but they don't differ in the nature of the prediction (*of*-genitive in both cases). Such kinds of scenarios are among the things that gave rise to the MuPDAR to be discussed here.

The MuPDAR approach essentially uses the logic of missing-data imputation to determine for every choice to (not) produce a certain unit by a learner what a native speaker would have chosen in the exact same linguistic and textual situation that the learner was in when he made the choice. That in turn allows the analyst to determine, first, whether learners make nativelike choices or not and, second, what the factors are that lead to learners making non-nativelike choices. The procedure would be based on the same kind of annotated as that in the previous section, but would then proceed as follows:

(1) apply some classifier (often (mixed-effects) regressions of the above type, but random forests have also been used and yet others would work as well) to the data of the native speakers to develop a model that 'learns' when native speakers do what (e.g. choose *of* vs. *s*-genitives) on the basis of all the variables mentioned above; determine whether that model 'does well' using classification and prediction accuracies, *C*-scores, R^2-scores, etc.

(2) if that model 'does well', use it to generate a native-speaker prediction for every data point in the learner data; this is the missing-data imputation step: we have

annotations for each learner choice and we have the learner choices, but what we need is missing, namely for every learner choice what a native speaker would have done and instead of asking several native-speaker annotators to provide that information, we 'ask' a statistical model to provide it instead;

(3) compare the actual learner choices to the imputed native speaker choices to determine (i) where they agree and (ii) where they disagree and capture the comparison in a variable; that variable can be binary (the learner made the native-speaker imputed choice or not) or numeric (how much does the learner's choice deviate from the native-speaker predicted probability?) and indeed other extensions are possible;

(4) run a second model (regression, random forests, …) with the binary or numeric variable that quantifies the learners' divergence from the native-speaker imputations as the dependent variable, and everything else annotated in the data as the independent variables and random effects as discussed above.

The interesting output is then the results of the second regression model – because they reveal what leads to non-native speakers making nativelike and non-native-like choices in what contexts, which is sometimes also facilitated by returning to the first classifier to see what native speakers do in what contexts. To see the power of this kind of analysis, the next section will give a brief overview of a recent MuP-DAR study (Wulff, Gries, & Lester, 2018), before Section 4 concludes.

3.2.3 *MupDAR: A very brief example*

The case study I want to briefly summarize to at least give a flavor of how many – not all – of the above things can come together is concerned with *that*-complementation in subject (see (5)), object (see (6)), and adjectival complementation (see (7)):

(5) a. The problem is that Jadzia likes Worf
b. The problem is Ø Jadzia likes Worf

(6) a. I thought that Julian likes Jadzia.
b. I thought Ø Julian likes Jadzia.

(7) a. I am glad that the Romulans entered the war
b. I am glad Ø the Romulans entered the war

This alternation has been well researched for English native speakers, but there is considerably less work on learner's choices of realizing, or not realizing *that* (; exceptions include Wulff et al. (2014) as well as Wulff (2016), on which Wulff, Gries, and Lester try to improve by (i) adding a psycholinguistic predictor, sur-

prisal (how surprising is the beginning of the complement clause given the verb of the main clause?), to the range of predictors and by (ii) using the MuPDAR approach. They had approximately 9.500 instances of *that*-complementation from native speakers of English as well as German and Spanish learners of English. They annotated all instances for a variety of features having to do with the lengths of the subjects of both main and complement clause, the lengths of material intervening between matrix clause subject and verb, matrix verb and complement clause, and the *that* slot and the complement clause. They also annotated for material preceding the main clause subject as well as the preference for or against *that* of the main clause verb and, as mentioned above, the degree of surprisal the beginning of the complement clause incurred after having seen the main clause verb.

Their analysis involved generalized linear mixed effects modeling with varying intercepts for subjects as well as varying intercepts for main clause verb forms nested into their lemmas. Their first regression on the native speakers indicated a good classificatory power (e.g., a *C*-score of 0.91) so they applied the fixed-effect coefficients of that model to the learner data. They then computed a *that*/Ø prediction for every learner example and computed a so-called deviation score from the predicted probabilities of nativelike choices that

- was zero when the learner made a nativelike choice;
- was >0 and ≤0.5 when the learner used *that* where the native speaker would not have;
- was <0 and ≥0.5 when the learner did not use *that* where the native speaker would have.

This deviation score was then the dependent variable of the second regression, a linear mixed-effects model with (i) all predictors as mentioned above, (ii) crucially, the predictor of L1 (German vs. Spanish) so as to determine whether speakers from different L1 backgrounds were affected differently by the predictors, and (iii) varying intercepts for speakers and forms nested into lemmas again. The final model was highly significant but accounted for a rather low degree of variance – however, several interesting effects emerged, three of which will be mentioned here to illustrate the different kinds of results and their fine resolution that this approach offers.

First, there was a highly significant main effect of surprisal, i.e. an effect that was observable across the board (surprisal did not participate in any interaction) and, thus, also across both learner L1s. Figure 7 shows that, as the first word of the complement clause becomes more surprising given the last word of the main clause, learners make significantly more nativelike choices. Both NS and NNS increase their complementizer use with higher rates of surprisal, and as before, the

NNS just do this with a higher overall baseline of *that*-use. This difference reflects the fact that even what is expected by native speakers remains rather unexpected to learners, a likely consequence of their lesser experience with naturalistic English use. Nevertheless, under conditions of high uncertainty, both groups appear to use *that* to smooth spikes in informational load (as reported for NS by Jaeger, 2010).

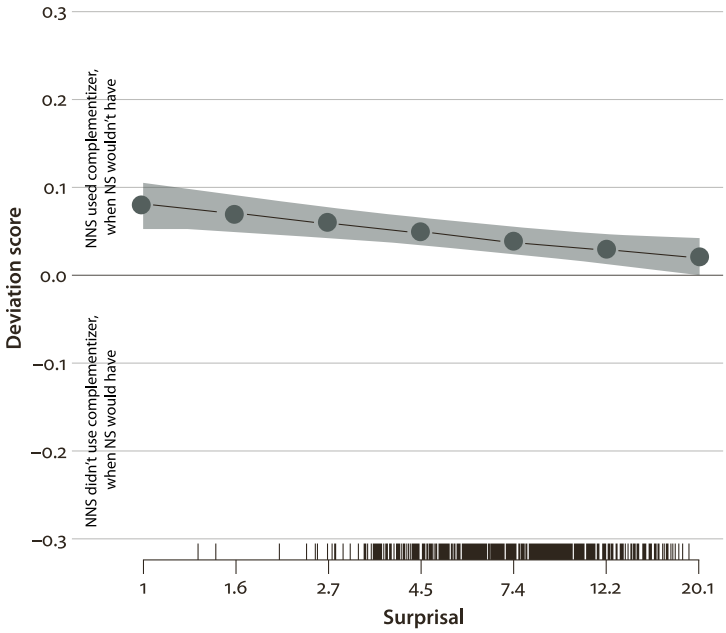

Figure 7. The effect of surprisal in regression 2 of Wulff, Gries, & Lester (2018)

Figure 8 is an example of an interaction: native speakers use *that* more in writing and less in speaking, but while the non-native speakers are fairly close to the native speakers in speaking, they still overuse the complementizer regardless of the length of the complement subject. In writing, on the other hand, the learners are more nativelike with longer subjects, but overuse *that* with short subjects (in particular *I*).

Finally, let's consider an interaction of a predictor with L1, i.e. an effect where the German learners differ from the Spanish learners. Figure 9 shows that, if there is material intervening between the subject and the verb of the main clause, then both German and Spanish speakers behave nativelike and use *that*, but when there is none, then both learner groups overuse *that*, but the Spanish speakers do so particularly much, which may be due to the fact that the Spanish analog to *that* is obligatory in all these kinds of complementation contexts.

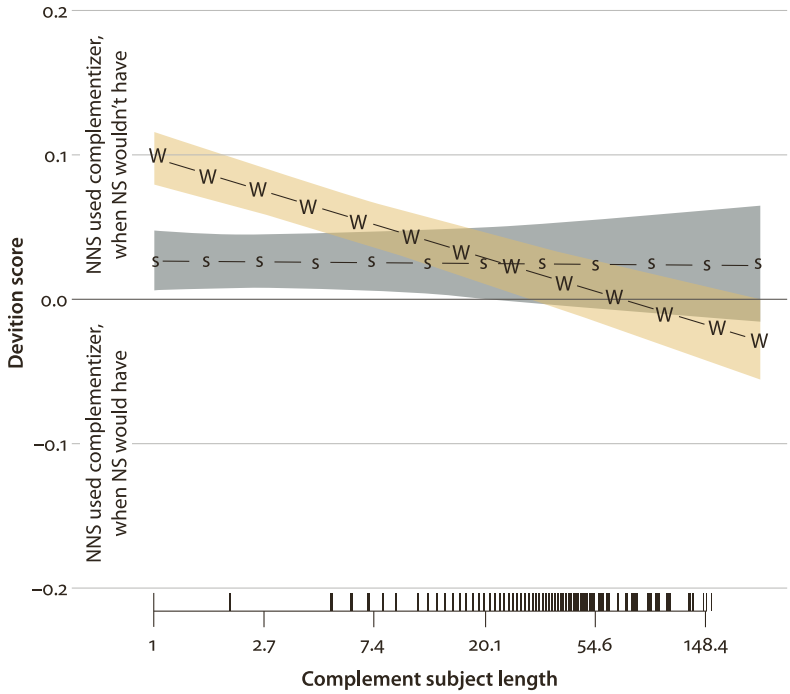

Figure 8. The effect of Mode : LengthComplementSubj in regression 2 of Wulff, Gries, & Lester (2018)

While this brief example could only scratch the surface of *that*-complementation, that was not its main point – its main point was to show how much more (than statistically flawed analyses studies of aggregate learner data) learner corpus research can do once sufficiently advanced methods are used.

4. Concluding remarks

To wrap up, let me reiterate the admittedly strong, but I believe supported, claims that I have made:

- over- and underuse studies on aggregate corpora are likely to be useless, and this is true for learner corpus research in particular but also corpus linguistics in general: the more frequent the phenomenon in question, the more anticonservative the currently still prevalent chi-squared/log-likelihood tests will be and the more of the actual variability in the data they will miss;
- since the degree of anticonservativeness is going to be phenomenon- and corpus-dependent, it is not even possible to just go back to previous results and

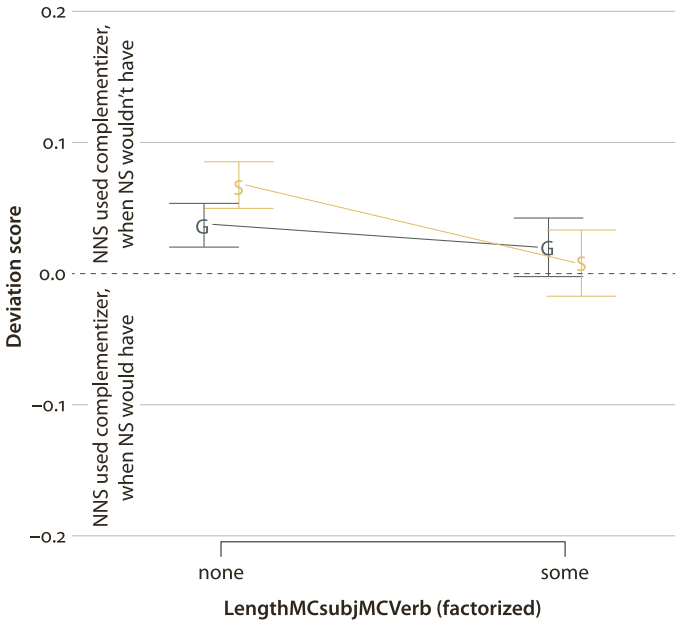

Figure 9. The effect of surprisal in regression 2 of Wulff, Gries, & Lester (2018)

'adjust' them in some sense: the degree to which they are wrong is unknowable; in addition, the methods employed in these previous studies violate statistical assumptions and suffer from a variety of other flaws discussed in many places and most recently/pertinently to learner corpus research in Paquot & Plonsky (2017);

- questions of over- and underuse *can be* studied in a statistically legitimate way for instance by generalized linear mixed-effects models or by linear models on data aggregated by speakers (the latter precludes the use of other predictors on cases, which makes it considerably less ideal and versatile); however,
- questions of over- and underuse *shouldn't be* studied like that because
- over- and underuse of any unit is a multifactorial issue, which means that such a monofactorial L1-as-the-only-predictor study cannot shed light on much and, even more importantly generally,
- to study even a truly *mono*factorial hypothesis, one needs a *multi*factorial test to avoid performing a test that (i) is overly anticonservative by pretending we know of nothing else that affects the phenomenon at hand and that (ii) does not even control the most elementary, general-purpose, and uncontroversial speaker- and text-specific predictors that are at work *everywhere*.

I then outlined ever so summarily one application that, while still not perfect, goes a long way in terms of addressing many if these issues. Now, without trying to be

polemic – just explicit – let me ask hopefully rhetorically: after having seen even this glimpse of the fine resolution the regression-based approach could provide on the *that*-complementation data (main effects, interactions of predictors, interactions with L1, repeated measures per speaker, verb form and lemma effects, effect sizes, confidence intervals/bands, …), do we really want to continue (i) analyzing such data by looking at no more than tables like Table 3, (ii) computing a variety of illicit chi-squared tests, and then (iii) simply concluding that learners overuse *that* and that, here, there seems to be no difference between German learners' and Spanish learners' use of *that* (59.5 and 60.2% respectively)? I hope not …

Table 3. Traditional input to an over-/underuse analysis of *that*-complementation

L1	**Ø**	***that***	**Totals (%)**
English	3214 (57%)	2420 (43%)	5634 (100%)
German	904 (40.5%)	1327 (59.5%)	2231 (100%)
Spanish	625 (39.8%)	945 (60.2%)	1570 (100%)
Totals (%)	**4743 (50.3%)**	**4692 (49.7%)**	**9435 (100%)**

And note that all these points of critique do not just apply to learner corpus research, but to much of corpus linguistics more broadly: As Paquot and Plonsky correctly observe, most of that field is held back by many of the same problems that learner corpus research is, although I do think that corpus linguistics in general is in a slightly better shape than the narrower field of learner corpus research.

This brings me to a final and more anecdotally motivated observation. For quite some time now, many learner corpus researchers seem to be a bit, let's say, miffed at the not-overly-warm reception of much of our work in the SLA community. My own feelings are ambivalent with regard to that, as one might guess from the main tenor of this paper and other work of mine. Yes, I do think observational data are much more useful than is often thought especially in experimental circles. This is not only because of the naturalness and, thus, hopefully the representativeness of the data and the generalizability of the findings, it's also because carefully balanced experimental designs expose subjects to a stimulus distribution that, because of its very balancedness, is quite different from the often very Zipfian distribution and highly intercorrelated structure. That in turn can easily lead to learning effects even over the course of just a few stimuli which, unless they are properly controlled, can affect analytical results (see Gries & Wulff, 2009; Doğruöz & Gries 2012; Gries to appear).

However, while the noisiness and the Zipfian distributions of observational data come with a higher degree of ecological validity, they also come with a higher degree of required statistical complexity: If the data are more natural and more

Table 4. Simplistic comparison of observational and experimental data

observational/corpus		experimental
low	artificiality/control	high
collinear & Zipfian	distribution	equal/balanced
harder	statistical analysis	simpler

noisy/messy, then it takes more careful and insightful and, typically, more complex statistical analysis to tease out from such data the real effects one is interested in. To an experimental SLA person, it will seem as much of learner corpus research implicitly criticizes them for controlled and thus unnatural data, but that, while we are doing better on the naturalness of the data, our statistical analyses don't even come close to doing justice to the complexity of everything they have already shown plays a role; they will think "often you people don't even distinguish between speakers or proficiency levels!". In other words, they might think "we're doing badly on the data side, but well on the analysis side, but you're doing well on the data side and badly on the analysis side". If learner corpus research wants to have the impact on SLA research that I think it deserves to have, we need to kick it up a notch: just having better data is not enough, the quantitative methods used need to evolve in tandem. While this paper has made strong claims, I hope they assist Paquot & Plonsky (2017) in bringing about the big methodological changes the field needs to be taken as seriously as we all want it to be taken.

Acknowledgements

This paper is a revised and extended version of a plenary talk I gave at the Learner Corpus Conference 2017 in Bolzano. I am grateful to Hilde Hasselgård and Magali Paquot for discussion of selected aspects of this paper, Sandra C. Deshors and Benedikt Heller for feedback on an earlier draft, and to Magali Paquot for access to Paquot & Plonsky (2017); the usual disclaimers apply.

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 55–76). Amsterdam: John Benjamins.

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 37–54). Amsterdam: John Benjamins.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). New York, NY: Springer.

Connor, U., Precht, K., & Upton, T. (2005). Business English: Learner data from Belgium, Finland, and the U.S. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 175–194). Amsterdam: John Benjamins.

Doğruöz, A. S., & Gries, S. Th. (2012). Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics*, 10(2), 401–426. https://doi.org/10.1075/rcl.10.2.07sez

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1–27. https://doi.org/10.18637/jss.v008.i15

Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 55–78). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.44.04gra

Gilquin, G., & Lefer, M.-A. (2017). Exploring word-formation in Learner Corpus Research: A case study on English negative affixes. Paper presented at the Learner Corpus Research conference 2017, Bolzano, Italy.

Gries, S. Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2), 109–151.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. Th. (2013). *Statistics for linguistics with R* (2nd rev. and ext. ed). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110307474

Gries, S. Th. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. https://doi.org/10.3366/cor.2015.0068

Gries, S. Th., & Adelman, A. S. (2014). Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms* (pp. 35–54). Cham: Springer.

Gries, S. Th., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, 9(1), 109–136. https://doi.org/10.3366/cor.2014.0053

Gries, S. Th. (to appear). Priming of syntactic alternations by learners of English: An analysis of sentence-completion and collostructional results.

Gries, S. Th., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, 7, 163–186. https://doi.org/10.1075/arcl.7.07gri

Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 33–61). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.45.06has

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6(2), 183–205. https://doi.org/10.1016/S1060-3743(97)90033-3

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Labov, W. (1982). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Neff van Aertselaer, J. & Bunce, C. (2012). The use of small corpora for tracing the development of academic literacies. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 63–83). Amsterdam: John Benjamins.

Paquot, M. & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

Wulff, S. (2016). A friendly conspiracy of input, L1, and processing demands: *that*-variation in German and Spanish learner language. In A. Tyler, L. Ortega, H. I. Park, & M. Uno (Eds.), *The usage-based study of language learning and multilingualism* (pp. 115–136). Washington, DC: Georgetown University Press.

Wulff, S., Lester, N. A. & Martinez-Garcia, M. M. (2014). *That*-variation in German and Spanish L2 English. *Language and Cognition*, 6(2), 271–299. https://doi.org/10.1017/langcog.2014.5

Address for correspondence

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
USA

stgries@linguistics.ucsb.edu