ICAME Journal • 48(1) • 2024
DOI: 10.2478/icame-2024-0002

sciendo

ICAME Journal

Against level-3-only analyses in corpus linguistics

Research Article

Stefan Th. Gries

University of California, Santa Barbara
and Justus-Liebig-Universität Giessen

Received 26 May 2023; accepted 14 August 2023

Abstract: In the last few decades, much work in corpus linguistics has attempted to discover, and then interpret, differences in the frequencies of use of linguistic elements (words, patterns, constructions, discourse features, etc.). It is probably fair to say that such studies were particularly frequent in (i) learner corpus research, (ii) corpus-based varieties research, and (iii) sociolinguistically motivated studies. For instance, many studies have discussed the differences in how often certain elements are used (i) in corpus data from native speakers vs. corpus data from learner from different L1 backgrounds, (ii) in corpora representing different inner- and outer-circle varieties, or (iii) by speakers in corpora representing people of different gender or sexual identities.

This paper will make the admittedly bold claim that any such study can in fact by definition unable to 'prove' what is often their main points, namely that the distributional differences found are in fact due to the one hypothesized explanatory variable(s) of L1, VARIETY, or, e.g., GENDER even when the distributional differences are significant and come with a decent effect size. To substantiate this claim, I will discuss some terminology from the family of methods known as multi-level modeling, namely the distinction between level-1, level-2, ... level-*n* variables and its relevance for many corpus studies. Second, I will then demonstrate how studies using only the above kinds of variables cannot distinguish the effect of their favored predictors from the effect of local/contextual level-1 variables. Third, in discussing this, I will exemplify how such effects need to be explored quantitatively instead.

Keywords: *learner corpus research • varieties research • regression • multi-level modeling • genitive alternation*

1 Introduction

Corpus linguistics is a distributional discipline: everything boils down to patterns observed in distributional data that we characterize via frequencies of occurrence of linguistic elements – in corpora, in files / by speakers, in contexts – and frequencies of co-occurrence of linguistic elements – with other linguistic elements or with a variety of contextual features. Such distributional data usually require quantitative methods, so it is not exactly surprising that (corpus) linguistics as well as usage-based linguistics has seen a strong quantitative turn (see Janda 2013, Jenset and McGillvray 2017, Joseph 2004, 2008, and the rise of textbooks introducing statistics to linguists). However, not only are there many statistical methods to choose from – even within one (family of) statistical method(s) – there is another probably even more important choice that needs to be made, namely what analytical level(s) of resolution to use for a specific study. In some sense, that might indeed be one of the most essential questions we face because our answer(s) determine(s) what the input will be to whatever quantitative method we will use and what the result will (be able to) tell us.

To explain the particular meaning of "analytical level(s) of resolution", let me use an example from alternation research, specifically the extremely well-known case of the genitive alternation exemplified

* *Corresponding author: Stefan Th. Gries E-mail: stgries@gmail.comc*

here in (1):

(1) a. the ambassador$_{possessor}$'s problem$_{possessed}$
b. the problem$_{possessed}$ of the ambassador$_{possessor}$

Now imagine I wake you up in the middle of the night, put the proverbial gun to your head, and ask you, "quick, what factors co-determine speakers' choices of constructions in the genitive alternation?" – what will you say? Even if you have never worked on the genitive alternation yourself, any kind of familiarity with alternation research would make it quite likely that you would mention factors such as

- the animacy of the possessor: *s*-genitives are more likely with human possessors;
- the length of the possessor: *s*-genitives are more likely with short possessors (esp. when the possessed is long);
- the discourse-givenness of the possessor: *s*-genitives are more likely with discourse-given possessors (esp. when the possessed is discourse-new);
- the length of the possessed: *of*-genitives are less likely with long possessors (esp. when the possessor is short);
- the discourse-givenness of the possessed: *of*-genitives are more likely with discourse-given possesseds (esp. when the possessor is new); etc.

(And, to foreshadow what is coming up below, the one answer that no one will give, ever, is "whether the speakers are German learners of English or not"). In a context where one is modeling the genitive alternation – i.e. where the constructional choice of *of* vs. *s* is a response variable that we might call `GENITIVE` – and in the language of multilevel modeling, we might call such variables **level-1 variables**: they are at the same 'data point'/observation level as the actual linguistic choice that speakers make when they go for one or the other construction. Because they describe the (immediate) linguistic/temporal context, they are the ones we are likely to think of first in the gun-to-your-head scenario and they have of course been the main target of probably thousands of studies.[1]

However, often other kinds of variables, variables at other, higher levels, are of main interest to a researcher. For instance, there is more and more interest in individual variation, either because that is actually what the main research question is about or because the advent of mixed-effects modeling has made all of us aware of how we need to control for how multiple different data points provided by a single speaker and/or, in an experimental context, to a certain stimulus 'belong together' or are correlated with each other. In a study of the genitive alternation, speaker *S021* might contribute four different *of*-genitives and two different *s*-genitives, each of which usually comes with different values on many level-1 variables because they involve (different combinations of) different lengths, givennesses, animacies of possessors and possesseds, etc. A speaker's level-1 constructional choices are typically all nested into the speaker, which means

- if I give you a genitive choice with all the context that would be required to annotate all factors (co-) determining the constructional choice, then, usually, only one speaker in your corpus data produced that one exact genitive with that specific context (see Section 3 for supporting evidence to that effect), so from the hypothetical I gave you, you can uniquely identify the speaker who produced exactly that one example, but
- if I give you the name of the speaker (*S021*), you cannot uniquely identify exactly one genitive example because the speaker produced, in our hypothetical, six different genitives.

Because of that, `SPEAKER` or, in many corpus-linguistic contexts, `FILE` can often be considered a **level-2 variable** and studies that specifically target individual variability in the genitive alternation would therefore target level-2 variables.

1 There are probably hundreds of studies of just the English dative alternation and the genitive alternation, and as soon as we add particle placement, the voice or future alternation, all relevant variationist sociolinguistic studies, and the studies on languages other than English, I doubt "thousands" is an exaggeration.

But there are still higher-level variables and much of the focus of this paper is on how they have been explored in many corpus-linguistic studies. For example, if your study of the genitive alternation was a learner corpus study, it is very likely that you have a variable such as L1 to consider: you might have native speaker (NS) data, where L1 is *English*, and non-native speakers, where the variable L1 has levels like, for instance, *Chinese* and *German*. In such data, the level-2 variable SPEAKER is nested into a level-3 variable L1 just like specific uses of genitives are nested into speakers: if you know the speaker, you know which level of L1 they come with, but if you know L1, you typically cannot uniquely infer one unique speaker from that because you have more than one speaker per level of L1.

Another example would be varieties research, where the situation is often similar or even more 'multi-leveled': one might have genitive choices and their contextual characteristics (level 1), which are nested into speakers/authors (level 2), which are nested into different newspapers that they write for and that were sampled (level 3), which are nested into different varieties of English (level 4), which are nested into variety types (inner- vs. outer- vs expanding circle) (level 5).

As a final example, consider how similar nesting kinds of situations might arise in sociolinguistic studies (with variables such as SEX or GENDER) or register studies (with variables such as formality/informality): if you know the speaker, you know the gender that speaker has in the corpus/data, but not the other way round.[2]

Given the above logic, in an alternation example like the present genitive case, variables such as L1, LANG, VARIETY, or SEX/GENDER might be level-3 variables, which have of course also been the target of many many studies in corpus linguistic research on, say, learner language, varieties, and sociolinguistic variation targeting precisely such level-3 variables. Against that background, the central point of this paper can be easily summarized in the following imperative: even if what you are mostly or exclusively interested in is a level-2 or a level-3 variable (according to the above definitions), do not do quantitative observational/corpus studies that include only level-2 or level-3 variables! I am fully aware that this imperative flies in the face of many corpus-linguistic studies that did nearly exactly that, but that is precisely the point/problem.

Learner corpus research, for instance, has seen very many studies discussing how learners (often from different L1s) over- or underuse certain linguistic elements (e.g., words or constructions). This kind of work established a veritable research tradition that may have started with Ringbom (1987) (see Cobb and Horst 2015: 187) before it was popularized in the 1990s; examples include, but are not limited to Granger and Tyson (1996), Hyland and Milton (1997), Altenberg and Granger (2001), Aijmer (2002), Altenberg (2002), Connor et al. (2005), Leńko-Szymańska (2008), Gilquin and Granger (2011), Hasselgård and Johansson (2011), Laufer and Waldman 2011, Min (2011), Neff van Aertselaer and Bunce (2012), Chen (2013), Tazegül (2015), Gilquin and Lefer (2017), Meriläinen (2017, 2020), Akutsu (2023), and many more.

Varieties research, too, has seen many studies that discuss how speakers from different varieties (often of English) use certain linguistic elements more or less than speakers of other varieties (e.g. Yeung 2009, Davydova et al. 2011, Bruckmaier 2017, or Parviainen and Fuchs 2019). Finally, sociolinguistic or register studies have often discussed differences in frequencies of linguistic elements between men and women, between adults and teenagers, between different socio-economic classes, between registers, between genders, etc. (e.g. Argamon et al. 2003, Cheshire 2007, Martínez 2011, or Alipour and Nooreddinmoosaa 2018, or most recently just a few weeks ago at QUALICO 2023, when a paper by Manning et al. stipulated lexical differences between different kinds of trans and cis-gendered speakers without any local/contextual level-1 variables).

Statistically speaking, such corpus-linguistic studies often use one of the two following 'statistical designs' (here exemplified once with a learner corpus example and once with a varieties example):

2 The way I am using level numbers here is maybe not the prototypical one in multilevel modeling, where "[the] lowest level (level 1) is usually defined by the individuals. However, this is not always the case. For instance, in longitudinal designs, repeated measures within individuals [as when one individual (author) provides multiple examples] are the lowest level. In such designs, the individuals are at level two, and groups are at level three" like I discussed here (see Hox, Moerbeek and van de Schoot 2018: 2).

```
ELEMENTyesvsno    ~ L1 # type-1 analysis
FREQofsomeelement ~ L1 # type-2 analysis
# or
ELEMENTyesvsno    ~ VARIETY # type-1 analysis
FREQofsomeelement ~ VARIETY # type-2 analysis
```

That means,

- in type-1 analyses, the study's response variable is individual lexical/constructional choices from a set of alternatives;
- in type-2 analyses the study's response variable is the frequency of lexical/constructional choices in some researcher-defined group of conditions/speakers/etc.

In a learner corpus context and with regard to this kind of approach, Cobb and Horst (2015: 188) argue that "the powers of simple frequency counts to shed light on learners' lexical development are considerable." I could not disagree more, however, and in this paper, which is conceptually a follow-up to Gries (2018), I will try to convince readers that the exact opposite is true, namely that such corpus studies, studies that focus on one or more level-2 variables only (i.e. variables that apply to speakers/writers) or on one or more level-3 variables only (such as L1/VARIETY/GENDER) are by definition virtually useless and need to stop. Yes, this is a harsh claim, but one that I think I will be able to substantiate in the following sections. In Section 2, I will walk readers through several 'type 1' kind of case studies of the genitive alternation in a learner corpus context with different resolutions (with two brief exceptions, all case studies are based on an actual, existing data set) to demonstrate why I have such a strong opinion about level-2/level-3-only studies. Then, in Section 3, I will follow this up briefly with a quick similar discussion regarding type-2 studies in a varieties context.

Two clarifying side remarks before we begin: first, the main example in the present paper (Section 2) is a learner corpus study, but the central target of this paper – *target* in the sense of what I am arguing against – is really not learner corpus research per se (which is why Section 3 continues my argumentation and exemplification with a varieties example). The central target is any kind of analysis that focuses on (frequencies of) occurrences of some linguistic expression that are studied without reference to level-1 context and level-2 variables on the basis of only comparing level-3 variables. Thus, learner corpus research here merely serves as a convenient didactic vehicle to make a more general point.

Second, this paper is also not denying that there is an increasing body of work that, as a reviewer put it, "gets it right" – that is indeed the case and that is a very welcome development; studies that succeed in discussing level-3 variables but correctly include lower-level variables in their quantitative design include, in learner corpus research, Lester (2019), Wulff and Gries (2021), or Dubois, Paquot and Szmrecsanyi (2022) and, in varieties research, Heller et al. (2017), many studies from current or former members of a research group in Leuven (e.g. Grafmiller and Szmrecsanyi 2018), or Gries et al. (2018), to name but a few. The point of this paper is that, while the number of studies that 'get it right' is increasing, the kind of study I am arguing against does not seem to 'die out', and one goal of this paper is to explain and exemplify irrefutably clearly why it should die out and to, thereby, accelerate that process.

2 The genitive alternation: A sequence of case studies

The genitive alternation data to be explored here are from Wulff and Gries (2021): nearly 3000 genitive choices from native speakers of English and non-native learners of English with a Chinese or German L1 background. The distribution of genitives across speakers is summarized here in Table 1 with absolute frequencies and the corresponding row percentages.

Table 1. Absolute/relative frequencies of genitives across the sample of three languages studied here.

	of-genitive	s-genitive	Sum
English	892 (89.6%)	104 (10.4%)	996
Chinese	872 (88.1%)	118 (11.9%)	990
German	817 (81.7%)	183 (18.3%)	1000
Sum	2581 (86.4%)	405 (13.6%)	2986

2.1 The problem: Level-3 variables only

We begin our exploration of the genitive alternation with what has been a very standard approach for many years: we look at the frequencies of the genitives and determine whether learners over- or underuse specific genitives. In much previous work, this has been done with one or more chi-squared tests or the conceptually very similar log-likelihood test. While it is often not clear from the published descriptions what statistical test was used exactly – one or more chi-squared tests for independence/difference or different chi-squared tests for goodness-of-fit – let's graciously assume here such a traditional learner corpus study did the latter, comparing each learner group against the native speakers with a separate chi-squared test for goodness-of-fit. If one did that more comprehensively than most previous studies actually did (by providing residuals as well as effect sizes for chi-squared tests for goodness-of-fit,[3] and by adjusting the *p*-values of those chi-squared tests for multiple comparisons, here with Holm's method), this would generate the following results:

```
         Chi.squared df      p Residual.for.of Residual.for.s effect.size
Chinese        2.311  1 0.1285          -0.491          1.439       0.000
German        66.034  1 0.0000          -2.626          7.690       0.015
```

And this is how much traditional work might have summarized this. There is a significant correlation between LANG and GENITIVE: the Chinese learners are not significantly different from the native speakers (X^2=2.31, *df*=1, *p*>0.12), but the German learners are (X^2=66.03, *df*=1, *p*<10-15); the way the German learners differ from the native speakers is an overuse of *s*-genitives (Pearson residual for *s*-genitives: 7.69), but the effect is extremely weak (effect size=0.015).

A slightly (!) better way would be to do this with a regression model, by using LANG as a predictor in a generalized linear model predicting GENITIVE. This approach is slightly better because it uses a more general and powerful method (that would also generalize to more complex scenarios) and it gives us an R^2-value as well as a straightforward evaluation of the classification/prediction performance of the model.[4] In this particular case, I first fit a null model with no predictors m.00 (for later) and then a model m.01, whose output shown here agrees with the chi-squared test in how the overall model is highly significant and how the Chinese learners are not significantly different from the native speakers (*p*=0.2964) whereas the German learners are (*p*<10–6):

```
$`Model summary`
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -2.1491     0.1036 -20.7417   0.0000
LANGchinese   0.1490     0.1427   1.0441   0.2964
LANGgerman    0.6529     0.1320   4.9464   0.0000
$`Significance of LANG`
          LRT Df     Pr..Chi.
LANG 28.79255  2 5.594692e-07
```

3 The effect size consists of relativizing the observed chi-squared value against the maximum possible one.

4 It also allows us to do other nice things such as the use of orthogonal contrast comparisons (something we will not pursue here) and a principled general way to do post-hoc tests.

```
$R2s
  McFadden R-squared Nagelkerke R-squared
         0.01214557          0.01751359
```

But it also confirms that, whatever overall significant effect there is, it is really very small: McFadden's R^2 and Nagelkerke's R^2 are tiny. Unsurprisingly (from the model's coefficients), this model therefore also comes with similarly miserable discriminatory/classification performance. It has zero discriminatory power and performs at baseline by always predicting of, as we can see in its 'confusion matrix':

```
     PRED
OBS    of
  of 2581
  s   405
```

We can visualize this result as in Figure 1: the *x*-axis represents the three L1s, the *y*-axis the predicted probabilities of *s*-genitives, and the points and error bars are predicted probabilities with their 95%-confidence intervals; post-hoc tests indicate that

- the Chinese learners use *s*-genitives more often than the native speakers, but not significantly so;
- the German learners use significantly more *s*-genitives than both the native speakers and the Chinese learners.

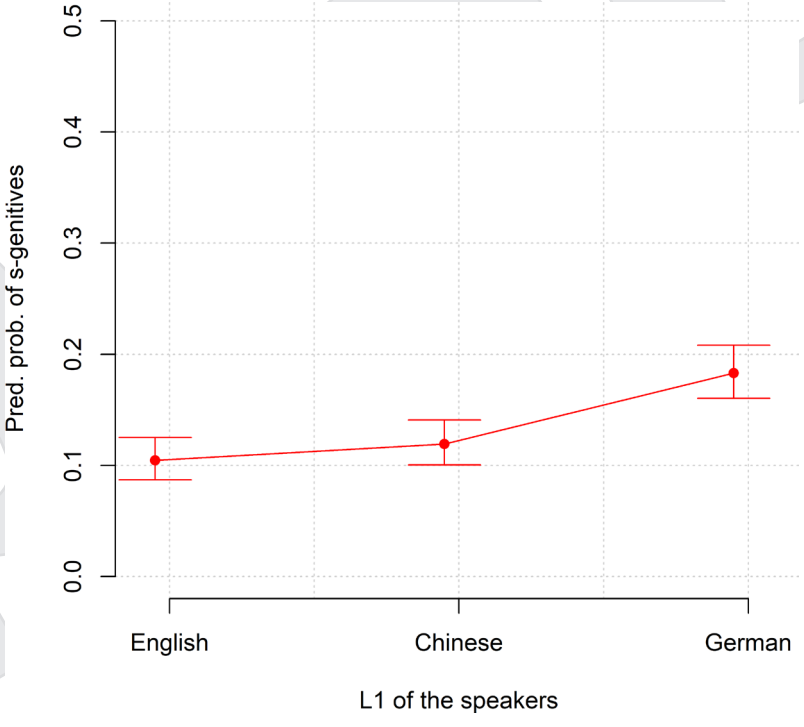

Figure 1. Predicted probabilities of *s*-genitives (analysis 1).

While the above analysis is already a bit better than many traditional studies – we have a more general/revealing effect size and we have confidence intervals, for instance – it is still useless and not just because the model happens to not come with a big R^2, but why? That is the topic of the next two sections.

2.2 Why is this supposedly "useless"? No level-2 variables

The first (and smaller) reason is based on the fact that the model uses only one predictor and, worse even, that that one predictor is a level-3 predictor and, thus, at a high level of abstraction. The problem that comes with this is that it by definition does not capture individual variation between speakers. Recall the distribution of the original data that we 'analyzed', shown here again for convenience:

Table 2. Absolute/relative frequencies of genitives across the sample of three languages studied here.

	of-genitive	s-genitive	Sum
English	892 (89.6%)	104 (10.4%)	996
Chinese	872 (88.1%)	118 (11.9%)	990
German	817 (81.7%)	183 (18.3%)	1000
Sum	2581 (86.4%)	405 (13.6%)	2986

From this cross-tabulation, which is all many traditional analyses considered, it is, however, completely unclear whether the data giving rise to the above summary table are distributed like this (I show data for only 6 Chinese and 6 German speakers in the rows and genitives in the columns) …

```
          of s
CNHK1001   2 0
CNHK1002   3 0
CNHK1003   7 0
CNHK1784   0 1
CNHK1785   0 1
CNHK1788   0 2
          of s
GEAU1001   1 0
GEAU1002   3 0
GEAU1003   5 0
GESA5043   0 1
GESA5044   0 2
GESA5045   0 1
```

… or like this:

```
          of s
CNHK1001   7 1
CNHK1002   8 1
CNHK1003   8 1
CNHK1105  14 2
CNHK1106  15 2
CNHK1109  16 2
          of s
GEAU1001   4 1
GEAU1002   5 1
GEAU1003   3 1
GEAU3073   3 1
GEAU3074   5 1
GEAU3075   4 1
```

In the two upper excerpts of a hypothetical version of the data, all speakers use only one kind of genitive and the overall totals result from the fact that the speakers who only produce *of*-genitives produce many more of those than speakers who only produce *s*-genitives produce *s*-genitives. In such

a case, the frequency table based on the level-3 variable would be completely misleading because it would fail to represent the complete either-or distribution that all learners exhibit. Put differently, nothing like the roughly 86.4% vs. 13.6% distribution of *of*-genitives vs. *s*-genitives is actually observed for even just a single speaker – they are all 100% vs. 0% or 0% vs. 100%. But in the two lower excerpts of another hypothetical version of the data, all speakers use both genitives in a proportion that resembles that of the overall distribution, and in such a case the frequency table based on the level-3 variable is nicely representative of each individual speaker.

Obviously, the two hypothetical scenarios are made up here precisely to represent these extremes. But just as obviously, without including any level-2 variable(s) in an analysis, one cannot see where on the cline between the two unnatural extremes one's actual data are: a proper analysis requires, among other things, that we include not just level-3 but also level-2 variables, something that many corpus-linguistic studies from all sorts of areas of application often do not do. So let's return to the real data and see what happens when we add a level-2 variable (as a random effect in a mixed-effects regression model m.02a) capturing speaker-specific baseline preferences for genitives:[5]

```
m.02a <- glmer(GENITIVE ~ 1 + LANG + (1|FILE), family=binomial, data=d)
$`Model summary`
            Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -2.4388     0.1583 -15.4042   0.0000
LANGchinese  -0.0587     0.1870  -0.3137   0.7537
LANGgerman    0.5156     0.1819   2.8344   0.0046
$`Significance of LANG`
          LRT npar     Pr.Chi.
LANG 13.47968    2 0.001182839
$`Model evaluation`
                VALUE
R2marginal     0.0141
R2conditional  0.3034
Simple McFadden 0.2659
C-score        0.9282
```

The model is very significant ($p \approx 0.0012$), but the effect of LANG seems very weak and most of the model's explanatory power comes from the varying intercepts, i.e. the level-2 variable with the speaker-specific variation: The $R^2_{marginal}$ of 0.0141 captures what the fixed effect is responsible for (i.e. very little) and $R^2_{conditional}$ of 0.3034 captures what all effects can account for, meaning the difference between the two is what the speaker-specific variability accounts for. In words, the level-2 variable SPEAKER accounts for 20 times as much of the deviance as the effect of LANG, and it is only because of that that the *C*-score is as good as it is.[6] Also, note that the coefficient for LANG: Chinese, which quantifies the log odds of *s*-genitives for Chinese learners compared to the native speakers, is negative. That means that, in this model, the Chinese learners have a *lower* predicted probability of *s*-genitives than the native speakers, not a *higher* one as in the previous model that did not account for speaker-specific variation.

If we generate the predictions that this model makes (using both fixed and random effects) and evaluate the confusion matrix, we see how bad this model still is: the model still nearly always predicts *of* – it only predicts *s* for a single German speaker (*GESA2010*).[7]

5 I am suppressing convergence warnings here in the output because such technicalities do not bear on the conceptual point here, the necessity of including level-2 variables in the first place. Also, here and further below, I restrict the output to the fixed-effects components and go with the simplest possible random-effects structure: varying intercepts for files/speakers. The "simple" McFadden R^2 is a heuristic comparison of the deviance of m.02a and the null model m.00 fitted earlier and is used here in addition to the more advanced R^2s for mixed-effects models (R^2_m and R^2_c) because it facilitates basic comparisons across different types of models.

6 If one makes predictions without the random effects, i.e. without the speaker-specific variability, the *C*-score becomes a terrible 0.5606856.

7 If one makes predictions without the random effects, the model is of course again even worse and always predicts *of* again.

```
$`Confusion matrix`
    PRED
OBS    of    s
  of 2579    2
  s   399    6
$Metrics
                         VALUE
Classification accuracy 0.8657
Precision for s         0.7500
Accuracy/recall for s   0.0148
Precision for of        0.8660
Accuracy/recall for of  0.9992
```

If we visualize this in a way that compares the present results with those from the previous section, we obtain Figure 2, and post-hoc tests indicate that

- the Chinese learners use *s*-genitives *less* often than the native speakers, but not significantly so;
- the German learners use significantly more *s*-genitives than both the native speakers and the Chinese learners.

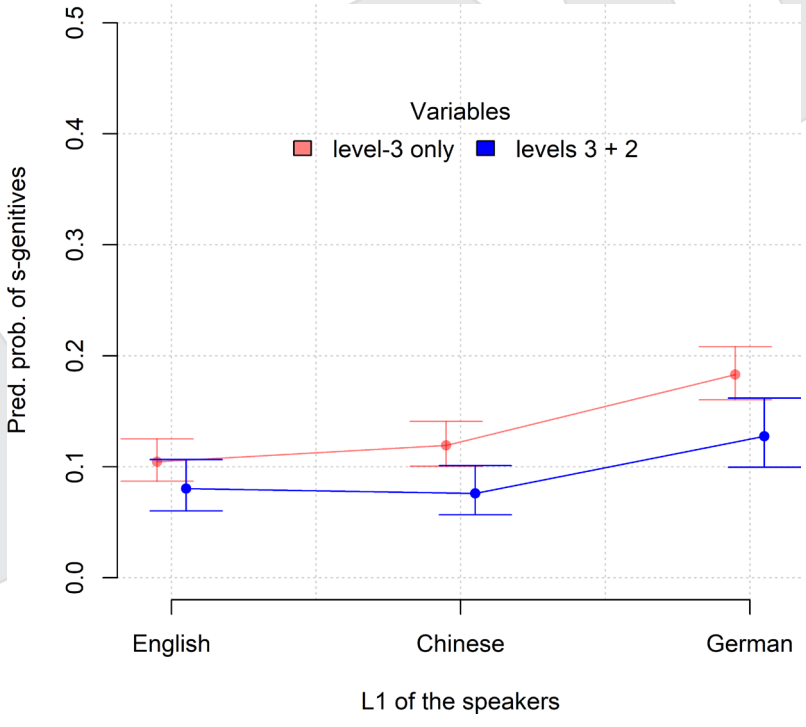

Figure 2. Predicted probabilities of *s*-genitives (analysis 2).

This model is statistically a tiny bit better than the first one, but it is still useless – why?

2.3 Why is this supposedly "useless"? No level-1 variables

The bigger reason why both our models with only the level-3 predictor or with the level-3 and level-2 variables are useless is that both embody an assumption that probably no linguist would ever subscribe to, that context does not matter – that is the conceptual equivalent of the statistical decision to not include contextual level-1 predictors. (An alternative implicit assumption just as unrealistic will be offered below.)

To appreciate that notion, let's briefly consider another hypothetical data set: it, too, distinguishes L1 speakers of English and L2 learners of English (but here I am not distinguishing different learner L1s just to keep matters simpler). And, this data set, like our real one, has more *of*-genitives than *s*-genitives. Finally, it has one level-1 predictor, namely a variable called LENGTHNPDIFF that quantifies the length difference between the possessor and the possessed, but – again for simplicity – as a factor with five levels which, with their label names, mean the same as in our original data set:

- if LENGTHNPDIFF=0, possessor and possessed are about equally long;
- if LENGTHNPDIFF>0 (or □0), the possessed is longer (or much longer) than the possessor, which should favor *s*-genitives.
- if LENGTHNPDIFF<0 (or □0), the possessor is longer (or much longer) than the possessed, which should favor *of*-genitives.

Here is the summary of these data:

```
  LANG     LENGTHNPDIFF GENITIVE
 eng:420   <<:120       of:559
 oth:500    <:150       s :361
            0:150
            >:200
           >>:300
```

Now, let's first do a traditional level-3 only LCR kind of study (but with a regression model); here is the cross-tabulation and the coefficients of a model m.02b:

```
addmargins(freq.tab.hyp3 <- table("LANG"=d.hyp3$LANG, "GENITIVE"=d.hyp3$GENITIVE))
     GENITIVE
LANG   of   s Sum
  eng 209 211 420
  oth 350 150 500
  Sum 559 361 920
summary(m.02b <- glm(
   GENITIVE ~ 1 + LANG,
   family=binomial,
   data=d.hyp3))$coefficients %>% round(4)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.0095     0.0976  0.0976   0.9223
LANGoth      -0.8568     0.1380 -6.2082   0.0000
```

There is a highly significant effect of LANG such that learners are much less likely to use *s*-genitives than native speakers. However, for this 'data', this interpretation is actually nonsense: the significant overuse of *of* by learners implied by the regression model is actually entirely due to LENGTHNPDIFF. The following tabulation shows there is a highly significant (X^2=346.64, *df*=4, *p*<10-73) and strong (Cramer's *V*=0.6138) relationship between LENGTHNPDIFF and GENITIVE:

```
             GENITIVE
LENGTHNPDIFF  of   s Sum
          -2 114  6 120
          -1 135  15 150
          0  120  30 150
          1  130  70 200
          2   60 240 300
          Sum 559 361 920
```

That means, if we add our level-3 predictor LANG, we see that, within each length relation, NS and NNS behave completely identically in terms of the percentage distribution of genitives. The 114 vs. 6 for when LENGTHNPDIFF is -2 become 19 vs. 1 (95% vs. 5%) for native speakers and 95 vs. 5 (i.e. also 95% vs. 5%) for learners, and so on:

```
                  GENITIIVE   of    s
LENGTHNPDIFF LANG
<<           eng            0.95 0.05
             oth            0.95 0.05
<            eng            0.90 0.10
             oth            0.90 0.10
0            eng            0.80 0.20
             oth            0.80 0.20
>            eng            0.65 0.35
             oth            0.65 0.35
>>           eng            0.20 0.80
             oth            0.20 0.80
```

Obviously, a regression model that 'does not know' about LENGTHNPDIFF is only too happy to overzealously attribute a lot of deviance to LANG, but a regression model that includes the level-1 predictor LENGTHNPDIFF and its interaction with the level-3 predictor LANG sees the irrelevance of LANG right away: all *p*-values involving LANG are 1 in both the full model and a final model after deleting the non-significant interaction LANG:LENGTHNPDIFF:

```
# a model w/ the interaction
summary(m.02b <- glm(
   GENITIVE ~ 1 + LANG*LENGTHNPDIFF,
   family=binomial,
   data=d.hyp3))$coefficients %>% round(4)
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.9444     1.0259 -2.8700   0.0041
LANGoth                 0.0000     1.1238  0.0000   1.0000
LENGTHNPDIFF-1          0.7472     1.1290  0.6618   0.5081
LENGTHNPDIFF0           1.5581     1.0851  1.4359   0.1510
LENGTHNPDIFF1           2.3254     1.0471  2.2207   0.0264
LENGTHNPDIFF2           4.3307     1.0410  4.1600   0.0000
LANGoth:LENGTHNPDIFF<   0.0000     1.2635  0.0000   1.0000
LANGoth:LENGTHNPDIFF0   0.0000     1.2044  0.0000   1.0000
LANGoth:LENGTHNPDIFF>   0.0000     1.1623  0.0000   1.0000
LANGoth:LENGTHNPDIFF>>  0.0000     1.1648  0.0000   1.0000
drop1(m.02b, test="Chisq") %>% data.frame %>% "["(-1, c(4,1,5))
                            LRT Df Pr..Chi
LANG:LENGTHNPDIFF 1.136868e-13  4       1
# a model w/out the interaction
summary(m.02b <- glm(
   GENITIVE ~ 1 + LANG+LENGTHNPDIFF,
   family=binomial, data=d.hyp3))$coefficients %>% round(4)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.9444     0.4446 -6.6221   0.0000
LANGoth          0.0000     0.1791  0.0000   1.0000
LENGTHNPDIFF<    0.7472     0.5004  1.4933   0.1354
```

```
LENGTHNPDIFF0    1.5581     0.4669  3.3373   0.0008
LENGTHNPDIFF>    2.3254     0.4483  5.1873   0.0000
LENGTHNPDIFF>>   4.3307     0.4520  9.5819   0.0000
drop1(m.02b, test="Chisq") %>% data.frame %>% "["(-1, c(4,1,5))
                  LRT Df      Pr..Chi
LANG            0.000  1 1.000000e+00
LENGTHNPDIFF 338.589  4 5.099511e-72
```

That means, in Section 2.2, we saw that a level-3 predictor alone can seem very highly significant but mask extremely different distributions with regard to level-2 variables – in this section, we see that a level-3 predictor alone can seem very highly significant but turn out to not do anything once a relevant level-1 predictor is also considered. And it is this combination of issues that renders analyses of unbalanced observational data with only level-2 and/or level-3 variables useless.

2.4 Improvement 2: Level-3, level-2, and level-1 variables

Let us now return to our actual data and add what every linguist would actually consider the normal kind of predictor, namely a small selection of level-1 variables; for now, we add them all as main effects (and here, LENGTHNPDIFF is in fact a numeric predictor):

```
summary(m.03a <- glmer(
   GENITIVE ~ 1 + LENGTHNPDIFF + POSSORANIM + POSSORNUMBER + LANG +
   (1|FILE),
   family=binomial, data=d),
   correlation=FALSE)$coefficients %>% round(4)
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept)             -0.0785     0.1872  -0.4192   0.6751
LENGTHNPDIFF             0.0685     0.0072   9.5040   0.0000
POSSORANIMinanimate     -3.0818     0.2162 -14.2565   0.0000
POSSORNUMBERplural      -4.3992     0.4514  -9.7464   0.0000
POSSORNUMBERirregplural -1.5422     0.2948  -5.2316   0.0000
LANGchinese              0.4842     0.2113   2.2915   0.0219
LANGgerman               0.2815     0.2062   1.3649   0.1723
```

The model is very significant, as we can see from comparing m.03a to a null model with the same random-effects structure (m.03null); I am shortening the output of anova a bit:

```
                              Chisq Df     Pr..Chisq
m.03a against a null model 667.2671  6 7.126818e-141
```

And note that, now, when contextual variables are controlled for, the German learners are not significantly different from the native speakers! But which variables are driving this model and how good is it?

```
$`Significance of predictors`
                     LRT npar      Pr.Chi.
LENGTHNPDIFF 124.395048    1 6.903563e-29
POSSORANIM   363.459950    1 4.968206e-81
POSSORNUMBER 247.319435    2 1.973636e-54
LANG           5.200966    2 7.423770e-02
$`Model evaluation`
                  VALUE
```

```
R2marginal      0.5556
R2conditional   0.6775
Simple McFadden 0.4749
C-score         0.9414
```

Interestingly, all effects but LANG are significant (!) and now most of the model's explanatory power comes from the fixed effects. And, for the first time, we have a better-looking confusion matrix:

```
$`Confusion matrix`
    PRED
OBS    of    s
  of 2524   57
  s   190  215
$Metrics
                          VALUE
Classification accuracy 0.9173
Precision for s         0.7904
Accuracy/recall for s   0.5309
Precision for of        0.9300
Accuracy/recall for of  0.9779
```

While LANG was not significant, let's explore its predictions in Figure 3; as we can see from post-hoc tests, the results indicate that, when level-1 and level-2 variables are controlled for, then, according to post-hoc tests,

- the Chinese learners use *s*-genitives significantly more often than the native speakers;
- the German learners use *s*-genitives more often than the native speakers, but not significantly so;
- the German learners use *s*-genitives less often than the Chinese learners, but not significantly so.

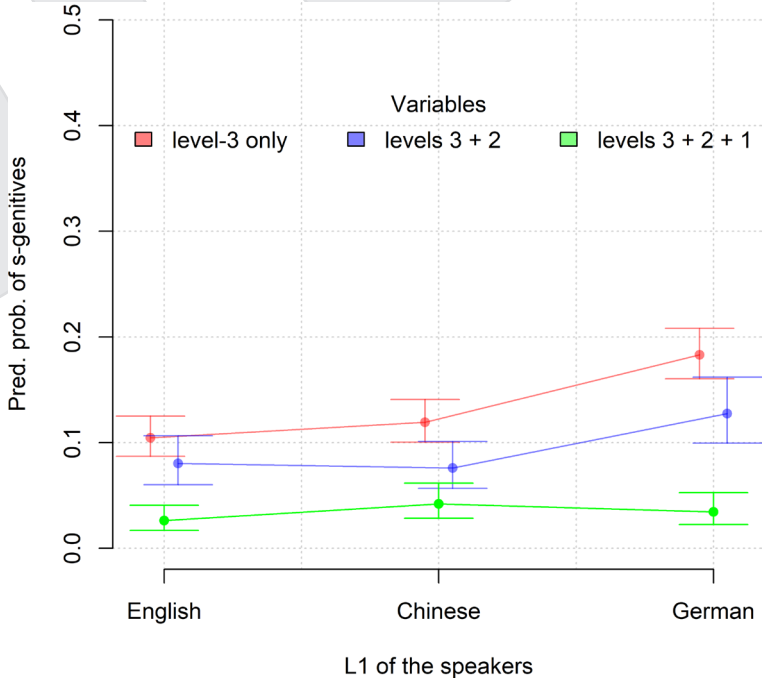

Figure 3. Predicted probabilities of *s*-genitives (analysis 3a).

While we now for the first time have a decent model, one that actually has some discriminatory power and a good fit as well as decent classification metrics, this is still not the best we should do as a learner corpus researcher (or a varieties researcher, etc.). This is because this model includes the level-3 predictor LANG only as a main effect, meaning it only checks whether speakers of different L1s are generally more or less likely to produce *s*-genitives – what it does not determine is whether any of the level-1 predictors works differently for speakers of different L1s, although that is probably what a learner corpus researcher would really want to know. That is what we turn to now.

2.5 Improvement 3: Level-3, level-2, and level-1 variables (w/ interactions)

The following is what, in a research context, one should have done right away: we use all variables we have, but let the level-3 predictor LANG interact with, here, two level-1 predictors, namely LENGTHNPDIFF and POSSORANIM:[8]

```
summary(m.03b <- glmer(
   GENITIVE ~ 1 + LANG * (POSSORANIM + LENGTHNPDIFF) + POSSORNUMBER +
   (1|FILE),
   family=binomial, data=d),
   correlation=FALSE)$coefficients %>% round(4)
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                       0.3414     0.2235   1.5277   0.1266
LANGchinese                      -1.0975     0.3459  -3.1732   0.0015
LANGgerman                       -0.0428     0.2755  -0.1553   0.8766
POSSORANIMinanimate              -3.9433     0.3466 -11.3763   0.0000
LENGTHNPDIFF                      0.0825     0.0151   5.4705   0.0000
POSSORNUMBERplural               -3.9972     0.4558  -8.7693   0.0000
POSSORNUMBERirregplural          -1.1885     0.2918  -4.0730   0.0000
LANGchinese:POSSORANIMinanimate   2.4236     0.4321   5.6093   0.0000
LANGgerman:POSSORANIMinanimate    0.6833     0.3935   1.7365   0.0825
LANGchinese:LENGTHNPDIFF          0.0015     0.0193   0.0757   0.9397
LANGgerman:LENGTHNPDIFF          -0.0302     0.0182  -1.6601   0.0969
```

But are all predictors significant? No, not all of them are:

- POSSORNUMBER is significant as a main effect across the three different L1s;
- the interaction LANG:POSSORANIM is significant, meaning the effect of POSSORANIM varies across the three L1s (even when basic speaker-specific variation is also controlled for in the model), but
- the interaction LANG:LENGTHNPDIFF is not significant, meaning whatever effect LENGTHNPDIFF has, it 'is the same' across the three different L1s.

```
list("Significance of predictors"=drop1(m.03b, test="Chisq") %>% data.frame %>%
"["(-1, c(3,1,4)))
$`Significance of predictors`
                          LRT npar      Pr.Chi.
POSSORNUMBER       189.92220    2 5.740082e-42
LANG:POSSORANIM     38.72897    2 3.891455e-09
LANG:LENGTHNPDIFF    4.88887    2 8.677514e-02
```

8 I do not include the interaction of LANG with POSSORNUMBER because of extremely low frequencies for the combination of *s*-genitives with plurals; German learners produce no *s*-genitives with regular plurals, Chinese learners produce only two. Also, for simplicity's sake and because this is not a mixed-effects modeling paper, I again omit convergence warnings and stick with the simplest of random-effect structures, the varying intercepts per file/speaker.

Let's therefore drop the non-significant interaction to arrive at our final model:

```
summary(m.03b <- glmer(
   GENITIVE ~ 1 + LANG*POSSORANIM + LENGTHNPDIFF + POSSORNUMBER +
   (1|FILE),
   family=binomial, data=d),
   correlation=FALSE)$coefficients %>% round(4)
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                       0.3684     0.2220   1.6597   0.0970
LANGchinese                      -1.1229     0.3422  -3.2817   0.0010
LANGgerman                       -0.0870     0.2748  -0.3165   0.7517
POSSORANIMinanimate              -3.9622     0.3411 -11.6167   0.0000
LENGTHNPDIFF                      0.0703     0.0073   9.6636   0.0000
POSSORNUMBERplural               -3.9023     0.4415  -8.8388   0.0000
POSSORNUMBERirregplural          -1.1508     0.2895  -3.9753   0.0001
LANGchinese:POSSORANIMinanimate   2.4892     0.4252   5.8545   0.0000
LANGgerman:POSSORANIMinanimate    0.6491     0.3844   1.6883   0.0914
```

The model is highly significant, as we can see from comparing m.03b to our null model with the same random-effects structure (m.03null):

```
                              Chisq Df    Pr..Chisq.
m.03b against a null model 707.0258  8 2.198099e-147
```

Also, it is good to see that most of the model's explanatory power comes from the fixed effects (see the R^2s) and the model's classificatory/predictive power is quite good again:

```
$`Confusion matrix`
     PRED
OBS    of    s
  of 2531   50
  s   189  216
$Metrics
                        VALUE
Classification accuracy 0.9200
Precision for s         0.8120
Accuracy/recall for s   0.5333
Precision for of        0.9305
Accuracy/recall for of  0.9806
$`Model evaluation`
                 VALUE
R2marginal      0.5581
R2conditional   0.6674
Simple McFadden 0.2742
C-score         0.9379
```

Let's see what the significant interaction of POSSORANIM:LANG means with some visualization and some post-hoc tests. In Figure 4, I avoid overcrowding the plot by now showing only the results for m.03b; the upper and lower lines represent the result for the animate possessors and the inanimate possessors respectively:

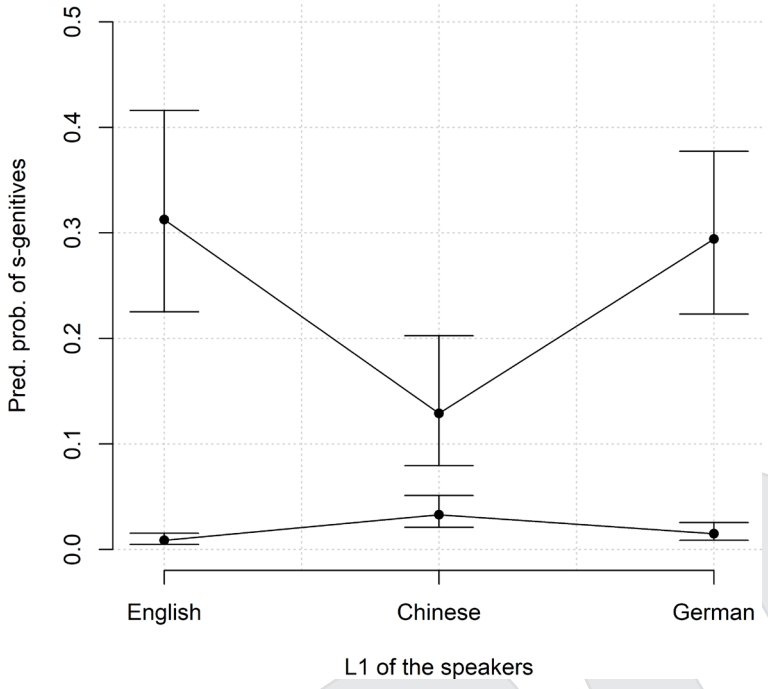

Figure 4. Predicted probabilities of *s*-genitives (analysis 3b).

The results for this model can be summarized as follows:

- there is a significant main effect of `POSSORNUMBER` (evaluation not shown here) such that the occurrence of *s*-genitives is highest for singular nouns (predicted probability: 0.084), followed by irregular plurals (pred. prob.: 0.028), followed by regular plurals (pred. prob.: 0.002) (all differences are highly significant in post-hoc tests and are not surprising);
- there is no significant interaction between `LENGTHNPDIFF:LANG`, which means speakers from all L1s behave alike with regard to the length difference between the NPs, and given the positive slopes the pattern we see there is a reliable short-before-long effect:
  - when the possessor is longer than the possessed, *s*-genitives become less likely;
  - when the possessor is shorter than the possessed, *s*-genitives become more likely;
- there is a significant interaction `POSSORANIM:LANG`: for each L1, animate possessors go with *s*-genitives more than inanimate possessors do, but they do so differently strongly:
  - for native speakers, the pred. prob. of *s*-genitives for animate possessors is 30.27% higher than for inanimate possessors ();
  - for Chinese learners, the pred. prob. of *s*-genitives for animate possessors is 9.74% higher than for inanimate possessors;
  - for Chinese learners, the pred. prob. of *s*-genitives for animate possessors is 28.03% higher than for inanimate possessors.

I hope we can agree that this is not only more 'correct' with regard to how the result has been arrived at statistically, but also much more interesting linguistically for any learner corpus researcher than a simplistic 'learners with L1 background *x* use more *s*-genitives than native speakers'.

2.6 Interim summary

It is instructive to once more compare all the different results because of how they highlight the risk of analyses that do not incorporate level-2 but especially level-1 predictors. **Analysis 1** used only the level-3 predictor of interest LANG and its result was that

- the Chinese learners are nativelike (in the sense of 'not significantly different from the native speakers') but use *s*-genitives a bit more than the native speakers;
- the German learners are not nativelike because of their massive overuse of *s*-genitives.

However, **analysis 2** indicated something different: once level-2 speaker-specific idiosyncrasies are controlled for,

- the Chinese learners are still nativelike, but use *s*-genitives a bit *less* than the native speakers;
- the German learners are again not nativelike because of their massive overuse of *s*-genitives.

But **analysis 3a** was again different: once level-2 speaker-specific idiosyncrasies and level-1 main effects are considered – and who would argue against that? – LANG is not significant anymore, meaning there is no difference in overall genitive preferences across all L1s at all. But since this analysis did not consider how LANG might affect the level-1 variables' effects, we fit **analysis 3b**, where we found that, when it comes to the effect of `POSSORANIM`,

- the Chinese learners are significantly not like the native speakers: they do use *s*-genitives significantly more with animate possessors than with inanimate possessors, meaning, they show the expected animacy effect, but that difference is, while significant, still fairly small (at least compared to the other two L1 speaker groups). In other words, for them animate possessors do not boost *s*-genitives that much;
- the German learners are nativelike: they use *s*-genitives significantly more with animate possessors than with inanimate possessors and they do so with predicted probabilities that are never significantly different from that of the native speakers (but significantly different from the Chinese learners).

This clearly indicates the dangers of analyses that include only level-3 predictors such as L1 or LANG, VARIETY, GENDER, …: even if we are ultimately interested in the effect(s) of such level-3 predictors – which learner corpus studies, many variety studies, many sociolinguistic studies, etc. of course are and legitimately so – to get at those level-3 predictors, we must not fit models that use only these predictors of interest and, thus, do neither control for speaker-specific level-2 effects nor for relevant contextual level-1 effects. Fitting such models amounts to pretending

- either that context does not matter – clearly an unrealistic assumption,
- or that speaker-specific variability is irrelevant *and* that all contextual variables in noisy Zipfian-distributed observational corpus data will on average be exactly the same or somehow magically balance each other out so they do not affect any other outcome.

I cannot imagine anyone who would seriously endorse either view. Yet, that is what all analyses quoted above that just used level-3 predictors implicitly endorsed. In other words, level-3-only analyses are implicitly based on assumptions no one would subscribe to ever, and the results they generate are not only extremely anticonservative – leaving all variability of the response variable 'up for grabs' to the level-3 predictor of interest makes it extremely more likely that that predictor will be significant in a statistical test – but it is also extremely not insightful: this is because, as we saw here, once we control for level-2 and level-1 variables, the actual effect of LANG is not a general main effect one of the rather dull 'speakers of L1 x do something more or less often' like 'Chinese learners use s-genitives more or less often than native speakers or other learners'. Instead, the real effect of LANG is much more interesting in how it does not modify the effect that the level-1 predictor LENGTHNPDIFF has – there, all learners behave nativelike – but the effect that another level-1 predictors has, namely POSSORANIM: Chinese learners are not nativelike because, to use the language of the Competition Model (Bates and MacWhinney 1989, Zhao and Fan 2021) they react less nativelike (here, much more weakly) to the cue 'animate possessor' than the German learners or the native speakers. Thus, the more fine-grained statistical resolution argued for here clearly affords us the possibility of much more fine-grained findings – which, on its own, is kind of self-evident – but the findings are now also more compatible with the important roles that everyone would agree individual variation and context play. Thus, my plea is that

we should just not have any level-3-only analyses of unbalanced observational data anymore: they are unrealistic and unreliable, period.

At the end of Section 1, I argued that level-3-only analyses usually amount to one of these two kinds of studies:

```
ELEMENTyesvsno    ~ L1 # type 1
FREQofsomeelement ~ L1 # type 2
```

This section discussed the first type, because our response variable was a binary genitive choice. In the following section, I will very briefly discuss a type-2 kind of example. I will do so with less detail (see Gries 2021: Exercises for Ch. 6 and Gries, under revision, for more details), but will still point out a few ways in which studies can face similar as well as additional problems as well.

3 Another brief example: Clause-final particles in three English varieties

The previous section discussed a kind of application where the response variable – whether in an actual regression context or in a more basic chi-squared/log-likelihood test content – is in fact the linguistic choice, e.g. one linguistic element (vs. not that element or another functionally similar element/construction) at what we called level 1. While we have seen that it is extremely problematic to run any such type-1 analysis with just level-2 and/or level-3 predictors, even such type-1 analyses have at least the advantage of targeting the response variable at the 'right' level of the individual observation (even if the predictors are at the 'wrong' levels). Unfortunately, the same cannot be said about what I above called type-2 analyses – why? To understand that, one has to see how type-2 analyses aggregate even more than type-1 analyses:

- type-1 analyses aggregate by counting every data point (i.e. instantiation of the response and some predictors) but by not distinguishing the data points on level 1 but only by level 3 and/or level 2;
- type 2 analyses aggregate even more namely by not even distinguishing all the data points anymore, but reducing them to frequency counts grouped according to level-3 and/or level-2 variables.

As an example, consider Parviainen and Fuchs (2019), a study on the use of *also* and *only* as clause-final particles in three different varieties of English. They do two separate but identical analyses, one on *also*, one on *only*, which is a first big mistake because they should have done one analysis on both particles at the same time and make PARTICLE a predictor that can interact with everything else. The point to be focused on here, however, is how they analyze their data. Consider Table 3, which shows the *also* part of their ICE-based data (from their appendix); the bold numbers are values I will refer to in a moment:

- `VARIETY`: the outer-circle varieties studied: *HKE* vs. *IndE* vs. *PhiE* from the relevant ICE components;
- `GENDER`: the genders of the speakers: *female* vs *male*;
- `AGE`: the age groups of the speakers: *14–25* vs. *26–35* vs. *36–50* vs. *>50*;
- `SPKRS`: the number of speakers in each group defined by `VARIETY`, `GENDER`, and `AGE`: between 0 and 170;
- `TOKENS`: the number of particle tokens found for each group defined by `VARIETY`, `GENDER`, and `AGE`: between 0 and 74;
- `WORDS`: the number of words found for each group defined by `VARIETY`, `GENDER`, and `AGE`: between 0 and 107,796;
- `TMWpaper`: the number of clause-final particle tokens (normalized to per million words) for all speakers in each of the 3×2×4=24 combinations of all levels of `VARIETY`, `GENDER`, and `AGE`: between 0 and 1705. (The formulation in the paper from which I inferred what I just described is this: "The final step in the analysis consisted of counting the number of tokens of clause-final *also* and *only* uttered by speakers of the different age and gender groups in each corpus, and counting the number of words contributed to the corpus by each of these groups." (p. 292).

Table 3. Parviainen and Fuchs's data for *also* as per their appendix 1.

VARIETY	GENDER	AGE	SPKRS	TOKENS	WORDS	TMWpaper
HKE	female	14–25	84	45	107,796	417
HKE	female	26–35	3	0	5304	0
HKE	female	36–50	6	0	5888	0
HKE	female	>50	0	0	0	0
HKE	male	14–25	15	5	16,599	301
HKE	male	26–35	2	0	2892	0
HKE	male	36–50	4	0	3204	0
HKE	male	>50	1	0	902	0
IndE	female	14–25	52	74	43,410	1705
IndE	female	26–35	33	33	29,456	1120
IndE	female	36–50	27	40	25,282	1582
IndE	female	>50	9	11	6757	1628
IndE	male	14–25	18	14	14,236	983
IndE	male	26–35	26	24	25,082	957
IndE	male	36–50	37	31	34,771	892
IndE	male	>50	37	35	31,927	1096
PhiE	female	14–25	170	37	93,197	397
PhiE	female	26–35	78	9	25,083	359
PhiE	female	36–50	54	6	11,674	514
PhiE	female	>50	0	2	3613	554
PhiE	male	14–25	69	6	32,386	185
PhiE	male	26–35	75	5	17,283	289
PhiE	male	36–50	106	6	6190	969
PhiE	male	>50	1	0	3937	0

The initial regression they fit on the *also* data shown in Table 3 would be written like this in R:[9]

```
m.also <- lm(TMWpaper ~ 1 + VARIETY*GENDER*AGE)
```

In other words, the response variable, here, is the relative frequency of clause-final *also*, which is what we called a type-2 analysis with a huge amount of aggregation: There

- are no level-1 predictors at all;
- are two level-2 predictors, i.e. variables at the level of speakers (a two-level GENDER and a 4-level variable AGE) that group every speaker into one of the resulting 2×4=8 level-2 variable groups, but the actual 907 speakers in their data are actually not distinguished (with a level-2 variable with 907 levels);
- is one level-3 predictor into which the speakers are nested (VARIETY), but, even worse, *all* the different instances of clause-final *also* for each combination of VARIETY, GENDER, and AGE are grouped/collapsed into a single relative frequency for that group.[10]

9 They say (p. 292) "Model selection was conducted using the step function, with *F*-tests as the selection criterion, and allowed for interactions of up to three variables". This is actually false, because the step function does not use *F*-tests but *AIC*. Also, this model can actually not even be fit because it is fully saturated. We have 24 combinations of predictor levels and 24 data points: fitting this model leads to non-computable standard errors, *t*-values, and *p*-values, and applying step to this model returns an error ("AIC is -infinity for this model, so 'step' cannot proceed").

10 It is also not clear how 0 female speakers of PhiE older than 50 could generate 2 clause-final *also*'s.

That means, on top of the absence of level-1 predictors, the model 'also does not know' that

- the frequency of 1705 clause-final *also*'s pmw by female IndE speakers between 14 and 25 years of age is based on 52 speakers;
- the very similar frequency of 1628 clause-final *also*'s pmw by female IndE speakers older than 50 is based on only 9 speakers, i.e only about $^1/_6$ of that number of speakers;
- the frequency of 983 clause-final *also*'s pmw by male IndE speakers between 14 and 25 years of age is based on 18 speakers;
- the very similar frequency of 969 clause-final *also*'s pmw by male PhiE speakers between 36 and 50 years of age is based on 106, i.e. nearly 6 times as many, speakers.

The amount of information loss that any such analysis incurs is hard to explain and yet also hard to overstate. Let me offer two illustrations to make it unmistakably clear what this kind of analysis does. First, we can apply the same reasoning here as in Section 2.2 above: using the first of the above four examples, the behavior of 52 different speakers (in likely very different Zipfian-distributed contexts) is condensed into just a single number, and from that number we do not even know whether the 74 tokens of clause-final *also* were produced by

- 1 speaker who produced all of them whereas 51 others did not produce any;
- 30 speakers producing one each and 22 speakers producing two each.

But the second illustration is maybe even clearer and for that we return to our genitive data. Let me remind the reader of what the only useful analysis of the data discussed here was, namely this generalized linear mixed-effects model:

```
glmer(GENITIVE ~ 1 +
   # four main effect predictors:
   LANG + POSSORANIM + LENGTHNPDIFF + POSSORNUMBER +
   # one interaction between two of them:
   LANG:POSSORANIM +
   # random effects (varying intercepts by speaker/file):
   (1|FILE), ...)
```

Consider now what it means to apply Parviainen and Fuchs's analytical approach to this data set. Since I only have access to AGE and GENDER for the vast majority of learners,[11] let's restrict 'the analysis' to the learners and, to approximate Parviainen and Fuchs's study even more, let's group AGE into four groups: 18–19 years (589 cases), 20–22 years (628 cases), 23–24 years (324 cases), 25–42 years (329 cases).

If we added AGEGROUP and GENDER to our genitive data/analysis (even just as main effects!), we could characterize the 'data space' of this otherwise proper mixed-effects modeling analysis as follows:

- right-hand side variables of minimally seven variables (plus potential interactions):
  - three level-1 predictors: POSSORANIM, LENGTHNPDIFF, and POSSORNUMBER;
  - three level-2 variables: AGEGROUP, GENDER, and the varying intercepts for 827 different speakers;
  - one level-3 variable: LANG;
  - plus whatever interactions between LANG and other predictors might be feasible;
- 1870 cases that instantiate 1070 different combinations of all predictors and the response.

But this only uses Parviainen and Fuchs's variables whereas we said we would apply Parviainen and Fuchs's analytical approach to this data, so what would that mean? They (i) only considered two level-2 predictors (what here would be AGEGROUP and GENDER) and one level-3 predictor (what here would be LANG) but no level-1 predictors and they (ii) conflated all speakers in each group. Thus,

11 For several German learners, values of AGE and GENDER were not available.

their approach to the data would actually be something like this: first, the big and intricate 'data space' from above would be reduced to something like Table 4 (where the new response variable GENRATIO is defined as the binary log of the proportion of *of*-genitives divided by the proportion of *s*-genitives):

Table 4. An equivalent to Parviainen and Fuchs's approach to our genitive data.

CASE	LANG	AGEGROUP	GENDER	GENRATIO
1	chinese	(17.5,19.5]	female	3.252387
2	german	(17.5,19.5]	female	2.473931
3	chinese	(19.5,22.5]	female	2.754888
4	german	(19.5,22.5]	female	2.390234
5	chinese	(22.5,24.5]	female	2.321928
6	german	(22.5,24.5]	female	1.984233
7	chinese	(24.5,45]	female	2.000000
8	german	(24.5,45]	female	1.594008
9	chinese	(17.5,19.5]	male	2.925999
10	german	(17.5,19.5]	male	NA
11	chinese	(19.5,22.5]	male	2.548893
12	german	(19.5,22.5]	male	3.201634
13	chinese	(22.5,24.5]	male	1.807355
14	german	(22.5,24.5]	male	1.559427
15	chinese	(24.5,45]	male	3.169925
16	german	(24.5,45]	male	2.925999

And then the statistical model would be something like this:

```
lm(GENRATIO ~ 1 +
   # level-2/3 predictors:
   LANG + AGE + GENDER +
   # their interactions
   LANG:AGE + LANG:GENDER + AGE:GENDER, ...)
```

In other words, their approach would mean that

1. instead of 1870 data points, we would have 16 (!);
2. instead of 1070 different usage conditions, we would have 16 (!);
3. instead of carefully separating 827 speakers who contribute differently many data points to the analysis (and get weighted accordingly) in a mixed-effects model, a normal linear model of the type Parviainen and Fuchs did, would treat every one of the only 16 data points the same because it would not even know that, for instance, the GENRATIO value of 2.754888 results from 112 female Chinese learners 20–22 years of age whereas the GENRATIO value of 2.321928 results from only 4 female Chinese learners 23–24 years old;
4. instead of controlling for contextual conditions, we would just ignore all of them, risking all the problems discussed in Section 2.3, Section 2.4, and Section 2.5:

a. the 'pretending context does not matter';
b. the correspondingly hugely inflated significance of level-2 and level-3 predictors; and
c. the inability to see how level-2 or level-3 predictors might really work.

Of course, someone might criticize my above argument by saying 'But Parviainen and Fuchs were not interested in POSSORANIM etc., they were only interested in a sociolinguistic apparent-time study, which is what motivated the inclusion of AGEGROUP and GENDER!' To which I would respond along the following lines: sure, I get that. But apart from the fact that point 3 above invalidates their kind of model(s) because of how the data points get weighted incorrectly, even an apparent-time study does not benefit from potentially hugely useless *p*-values. Why would the *p*-values be hugely useless? Because they could be (i) hugely anticonservative, because their modeling allows all variability in the data to be claimed by just the level-2/-3 predictors (as in Section 2.1), or they could be (ii) hugely conservative, because their approach reduces the sample size by two orders of magnitude. This problem cannot be 'solved' by ignoring context or hoping that contexts will somehow be magically comparable enough over the different age/gender/language slices of the data – that is what creates the problem in the first place. In no statistical, or linguistic, universe is reducing the richness of the 'data space' the way these analyses do a good idea.

4 Concluding remarks

The conclusions are, I believe, straightforward and were in fact already spelled out at the beginning in the imperative, repeated here: even if what you are mostly or exclusively interested in is a level-2 or a level-3 variable (according to the above definitions), do not do quantitative observational/corpus studies that include only level-2 or level-3 variables. Type-1 and type-2 studies that only use level-2 variables (file/speaker-specific variables) and/or only even higher-level variables like L1 or VARIETY without controlling for local/contextual variables run the risk of hugely exaggerating the role of the level-2/-3 variables of interest while at the same time completely missing how any of these variables supposedly of interest interact with local context. In learner corpus research, how can one seriously try and explain over-/underuse of something by learners with transfer if true linguistic/contextual parameters were not even considered in the quantitative evaluation?! In corpus-based varieties research, how can one seriously try and explain over-/underuse of something by speakers of one variety with their evolutionary distance to a historical source variety if true linguistic/contextual parameters were not even considered in the quantitative evaluation?! This has two quite brutal-sounding consequences but I honestly see no way around them:

- all studies cited above that use one of these level-2 and/or level-3 protocol could be wrong, completely wrong in fact, simply because they report results of the type reported here for the genitives in Section 2.1, which we have seen are utterly false or at least incomplete. If these were studies on something truly important, like the efficacy or safety of a vaccine, no one would accept their conclusions and they would all need to be done again – all of them.
- there is virtually no place for simple chi-squared-/log-likelihood-based frequency comparisons in learner corpus studies or variety studies or any field with comparable questions and methods, because they pretty much by definition do not permit the inclusion of the multiple levels of analysis that are required to avoid the pitfalls discussed in Section 2.1, Section 2.2, and Section 2.3. Even the seemingly simplest analysis concerned with anything like over-/underuse requires some kind of modeling approach that can handle multiple levels of analysis that corpus data present by definition.

Again, I know these are harsh conclusions, but the results of Section 2.1 and the subsequent reanalyses are clear. While I do not expect the field to actually redo all those studies like we should, if this paper makes sure that no more level-2/-3-only analyses are done ever again, I will be satisfied with its impact. In this context, one question posed by a reviewer is worth addressing here, namely the question of how my imperative applies to cases not involving alternations where the inclusion of local/contextual level-1 variables is fairly straightforward.

One part of the response to this question has already been mentioned above because the Parviainen and Fuchs study is not an alternation study: they do not study 'clause-final *also/only* vs.

non-clause-final *also/only*' but without level-1 variables. In their case, the solution could actually be that they 're-conceptualize' their study as an alternation study: they could look at, say, *also* and make the response not what it is now (the frequency of *also* in some combinations of not-level-1 predictors) but something like is CLAUSEFINALalso: *no* vs. *yes*. Then, we can obviously include local/contextual level-1 predictors. Another example is a study that I critiqued in the conceptual precursor to this paper, Gries (2018), namely Hasselgård and Johansson's (2011) study of *quite*. Their study, too, is not an alternation study because they do not study '*quite* vs. something else' but their paper exhibits the same no-level-1 variables kind of problem. In their case, my (2018) study shows how at least level-2 can immediately be added by treating every word in every file of the corpus (level-2) as an instance of a then binary response like WORDisQUITE: *no* vs. *yes*, which already makes the results quite a bit better because of the massive disaggregation of data points that step entails. A similar logic applies in cases where the focus of the study is a construction – either a fully schematic very general construction (like passives), a schematic construction with more specific semantics (like the ditransitive or other argument structure constructions), or partially lexically-filled constructions. In each case, one could either identify and annotate instances of an alternating construction or, minimally, if one does not want to annotate instances of an alternating construction, one could identify local/contextual descriptors of the construction in question and see whether, when those are included in a quantitative description, they already account for the use of the construction in question – if that was the case, Occam's razor would seem to lead to the conclusion that recourse to level-2 or level-3 variables is not necessarily supported. Thus, much like I mostly used learner corpus research as a vehicle for my recommendations, but with conclusions applying to many other fields with similar studies, I would like to think that, while my main exemplification used an alternation study as a vehicle, the overall conclusions do still apply and I would be the first to welcome any suggestions of how to port them to cases where the integration of local/contextual variables is, or at first sight seems, less straightforward because the alternative – continuing to focus on level-3 only while ignoring level-2 and level-1 variables – is not an option, that much I hope we can agree on already.

References

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

Akutsu, Sumie. 2023. The use of university students' English essays and reflection comments to provide more effective feedback. In U. Widiati et al. (eds.). *Proceedings of the 20th AsiaTEFL-68th TEFLIN-5th iNELTAL Conference (ASIATEFL 2022)*, 697–707. (https://doi.org/10.2991/978-2-38476-054-1_60)

Alipour, Mohammad and Mona Nooreddinmoosaa. 2018. Informality in applied linguistics research: Comparing native and non-native writings. *Eurasian Journal of Applied Linguistics* 4 (2): 349–373.

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

Altenberg, Bengt and Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22 (2): 173–195.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk* 23 (3): 321–346.

Bates, Elizabeth and Brian MacWhinney. 1989. Functionalism and the competition model. In B. MacWhinney and E. Bates (eds.). *The crosslinguistic study of sentence processing*, 3–73. New York, NY: Cambridge University Press.

Bruckmaier, Elisabeth. 2017. *Getting at GET in World Englishes*. Berlin and Boston: de Gruyter.

Chen, Meilin. 2013. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics* 18 (3): 418–442.

Cheshire, Jenny. 2007. Discourse variation, grammaticalisation and stuff like that. *Journal of Sociolinguistics* 11 (2): 155–193.

Cobb, Tom and Marlise Horst. 2015. Learner corpora and lexis. In S. Granger, G. Gilquin and F. Meunier (eds.). *The Cambridge handbook of learner corpus research*, 185–206. Cambridge: Cambridge University Press.

Connor, Ulla, Kristen Precht and Thomas A. Upton. 2002. Business English: Learner data from Belgium, Finland, and the U.S. In S. Granger, J. Hung and S. Petch-Tyson (eds.). *Computer learner corpora, second language acquisition, and foreign language teaching*, 175–194. Amsterdam and Philadelphia: John Benjamins.

Davydova, Julia, Michaela Hilbert, Lukas Pietsch and Peter Siemund. 2011. Comparing varieties of English: Problems and perspectives. In P. Siemund (ed.). *Linguistic universals and language variation*, 291–324. Boston and Berlin: De Gruyter Mouton.

Dubois, Tanguy, Magali Paquot and Benedikt Szmrecsanyi. 2022. Alternation phenomena and language proficiency: The genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory*. Ahead-of-print: https://doi.org/10.1515/cllt-2021-0078.

Gilquin, Gaëtanelle and Sylviane Granger. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee and M. Hundt (eds.). *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 55–78. Amsterdam and Philadelphia: John Benjamins.

Gilquin, Gaëtanelle and Marie-Aude Lefer. 2017. Exploring word-formation in learner corpus research: A case study on English negative affixes. Paper presented at Learner Corpus Research 2017, Bolzano, Italy.

Grafmiller, Jason and Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world. A study in comparative sociolinguistic analysis. *Language Variation and Change* 30 (3): 385–412.

Granger, Sylviane and Stephanie Tyson. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15 (1): 17–27.

Gries, Stefan Th. 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 Th. 2021. *Statistics for linguistics with R*. 3rd rev. and ext. ed. Boston and Berlin: De Gruyter.

Gries, Stefan Th. 2023. Corpus-linguistic and computational methods for analyzing communicative competence: Contributions from usage-based approaches. In M. H. Kanwit and M. Solon (eds.). *Communicative competence in a second language: Theory, method, and applications*, 115–131. New York and London: Routledge.

Gries, Stefan Th. under review. On proper regression modeling in varieties research: A critique, with advice for the future.

Gries, Stefan Th., Tobias J. Bernaisch and Benedikt Heller. 2018. A corpus-linguistic account of the history of the genitive alternation in Singapore English. In S. C. Deshors (ed.). *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, 245–279. Amsterdam and Philadelphia: John Benjamins.

Hasselgård, Hilde and Stig Johansson. 2011. Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds.). *A taste for corpora: In honour of Sylviane Granger*, 33–61. Amsterdam and Philadelphia: John Benjamins.

Heller, Benedikt, Benedikt Szmrecsanyi and Jason Grafmiller. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45 (1): 3–27.

Hox, Joop J., Mirjam Moerbeek and Rens van de Schoot. 2018. *Multilevel analysis: Techniques and applications*. 3rd ed. New York and London: Routledge.

Hyland, Ken and John Milton. 1997. Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6 (2): 183–205.

Janda, Laura A. (ed.). 2013. *Cognitive linguistics: The quantitative turn*. Boston and Berlin: Mouton de Gruyter.

Jenset, Gard B. and Barbara McGillvray. 2017. *Quantitative historical linguistics*. Oxford: Oxford University Press.

Joseph, Brian D. 2004. On change in *Language* and change in language. *Language* 80 (3): 381–383.

Joseph, Brian D. 2008. Last scene of all … *Language* 84 (4): 686–690.

Laufer, Batia and Tina Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61 (2): 647–672.

Leńko-Szymańska, Agnieszka. 2008. Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Aile* 27(2008), retrieved from https://journals.openedition.org/aile/4213, 23 May 2023.

Lester, Nicholas A. 2019. *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research* 5 (1): 1–32.

Manning, Theodore, Eugenia Lukin, Ross Klein and Patrick Juola. 2023. Construction and analysis of a map-based corpus for tracking linguistic variation and demographic characteristic identification. Paper presented at QUALICO 2023, University of Lausanne, 28 June 2023.

Martínez, Ignacio M. Palacios. 2011. "I might, I might go I mean it depends on money things and stuff": A preliminary analysis of general extenders in British teenagers' discourse. *Journal of Pragmatics* 43 (9): 2452–2470.

Meriläinen, Lea. 2017. The progressive form in learner Englishes: Examining variation across corpora. *World Englishes* 36 (4): 760–783.

Meriläinen, Lea. 2020. The interplay between universal processes and cross-linguistic influence in the light of learner corpus data: Examining shared features of non-native Englishes. In B. Le Bruyn and M. Paquot (eds.). *Learner Corpus Research meets Second Language Acquisition*, 67–95. Cambridge: Cambridge University Press.

Min, Sujung. 2011. A corpus-based analysis of EFL learners' use of discourse markers in cross-cultural communication. *English Language and Literature Teaching* 17 (3): 177–194.

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

Parviainen, Hanna and Robert Fuchs. 2019. 'I don't get time only': An apparent-time investigation of clause-final focus particles in Asian Englishes. *Asian Englishes* 21 (3): 285–304.

Péry-Woodley, Marie-Paule. 1990. Contrasting discourses: Contrastive analysis and a discourse approach to writing. *Language Teaching* 24 (3): 205–214.

Ringbom, Håkan. 1987. *The role of the first language in foreign language learning*. Bristol: Multilingual Matters.

Tazegül, Assiye Burgucu. 2015. Use, misuse and overuse of 'on the other hand': A corpus study comparing English of native speakers and learners. *International Online Journal of Education and Teaching* 2 (2): 53–66.

Wulff, Stefanie and Stefan Th. Gries. 2021. Explaining individual variation in learner corpus research: Some methodological suggestions. In B. Le Bruyn and M. Paquot (eds.). *Learner corpora and second language acquisition research*, 191–213. Cambridge: Cambridge University Press.

Yeung, Lorrita. 2009. Use and misuse of 'besides': A corpus study comparing native speakers' and learners' English. *System* 37 (2): 330–342.

Zhao, Helen and Jason Fan. 2021. Modeling input factors in second language acquisition of the English article construction. *Frontiers in Psychology* 12 (653258).