Received: 18 February 2024 | Revised: 27 May 2024 | Accepted: 24 June 2024

DOI: 10.1111/weng.12694

WORLD ENGLISHES WILEY

ORIGINAL ARTICLE

On regression modeling in varieties research

Stefan Th. Gries[1,2]

[1] Department of Linguistics, UC Santa Barbara, Santa Barbara, California, USA

[2] Department of English, JLU Giessen, Giessen, Germany

Correspondence
Stefan Th. Gries, Department of Linguistics, University of California, Santa Barbara, Santa Barbara, CA 93106-3100, USA.
Email: stgries@linguistics.ucsb.edu

Abstract

One particularly prominent methodological development in linguistics is what has been termed the "quantitative turn": Not only are more and more studies using statistical tools to explore data and to test hypotheses, the complexity of the statistical methods employed is growing as well. This development is particularly prominent in all kinds of corpus-linguistic studies: 20 years ago chi-squared tests, *t*-tests, and Pearson's *r* reigned supreme, but now more and more corpus studies are using multivariate exploratory tools and, for hypothesis testing, multifactorial predictive modeling techniques, in particular regression models (and, increasingly, tree-based methods). However welcome this development is, it, and especially its pace as well as the fact that few places offer rigorous training in statistical methods, comes with its own risks, chief among them that analytical methods are misapplied, which can lead imprecise, incomplete, or wrong analyses. In this paper, I will revisit a recent regression-analytic study in the research area of English varieties (on clause-final *also* and *only* in three Asian Englishes) to:

- highlight in particular three fundamental yet frequent mistakes that it exemplifies;
- discuss why and how each of these mistakes should be addressed;
- reanalyze the data (as far as is possible with what is available) and show briefly how that affects the analysis's results and interpretation.

1 | WHAT IS EXPLORED AND HOW?

The use of more advanced quantitative methods has seen a steady increase in corpus linguistics in general and in corpus-based studies of varieties (Parviainen & Fuchs, 2019). Although much of corpus linguistics was long dominated by statistical tests that were monofactorial in nature—chi-squared tests, *t*-tests, simple monofactorial correlations like Pearson's *r* and Spearman's *ρ*—now papers routinely use multifactorial regression, classification and regression trees, random forests, and other modeling techniques. However, welcome this development is in general, the adoption of these methods also comes with a greater responsibility of the researcher to apply them properly so as to both harness the increased power and versatility of such methods and avoid potential pitfalls they can pose. These three things—the methods' power, their pitfalls, and the corresponding demands these pose for researchers—are at the heart of this methodological paper, in which will exemplify several of these things using a corpus-based varieties study published a few years ago.[1] That study is Parviainen and Fuchs (2019; henceforth P&F), a corpus-based study of the two focus particles *also* and *only* and their frequency of clause-final use in the International Corpus of English (ICE) components for India (ICE-IND), Hong Kong (ICE-HK), and the Philippines (ICE-PHI); two examples for this kind of use are the following (P&F, p. 285):

(1) a. I do not have to work also (ICE-HK:S1A-004#967).
b. I do not get time only (ICE-IND:S1A-052#37)1.

This clause-final positioning of the particles is possible, but dispreferred, in Inner Circle varieties, which generally prefer a clause-medial position between subject and verb, but the clause-final position is attested in many Asian Englishes and particularly frequent in Indian English (IndE). P&F are interested in the ultimately epicentral question of whether "IndE—the variety where clause-final 'also' and 'only' are used most frequently—could have contributed to the emergence of the feature in SinE [Singaporean English], PhiE [Philipppine English] and HKE [Hong Kong English]" (p. 286). After ruling out an explanation of these particles' distributions in HKE and PhiE based on substrates and given the absence of diachronic spoken corpora covering the varieties of interest, P&F adopt the sociolinguistic apparent-time method, according to which "speakers from different age groups can be argued to represent the type of language that was used in their adolescence" (p. 290). Specifically, they state that:

> By examining the age and gender of those who use innovative 'also' and 'only' in ICE-IND, ICE-PHI and ICE-HK, it is possible to estimate how established the feature is in each variety and, consequently, provide further evidence on the potential influence of IndE in the Southeast Asian region. If clause-final 'also' and 'only' are used more frequently by younger speakers than by older speakers and/or more frequently by female speakers than by male speakers, we can infer that these features are becoming more frequent in these varieties. (P&F, p. 286)

Their methods section articulates the following expectations:

> Therefore, if the relative proportion of the use of innovative 'also' and 'only' by young (and) female speakers is higher than that of older (and) male speakers in both PhiE and HKE (when compared with IndE), this would indicate that the use of the feature is a more recent innovation in the former two varieties. This, in turn, would lend further support to the argument that IndE is an emerging epicentre in South(east) Asia. (P&F, p. 291)

> [we wanted] to determine: whether younger and/or female speakers are more likely to use the feature in question; and whether the three varieties differ in this regard as well as in the overall frequencies of clause-final 'only' and 'also'. (p. 292)

Finally, in their results section, they say:

> Specifically, we expected to find that younger speakers use these [clause-final particles] more often than older speakers, and that female speakers use them more often than male speakers. (p. 294)

Their analyses are based on the private-conversation parts of the studied corpus components, for which relevant metadata for the speakers are available. They retrieved all instances of clause-final *also/only* from these corpus parts and put together a data frame shown here as Table 1 (their appendix 1 with two cosmetic changes and my bolding is explained further below).

These are the contents of each column of this table:

- CASE: A unique identifier for each row (I added this to each row, P&F do not provide this strictly speaking unnecessary variable).
- PARTICLE: The clause-final particles studied: *also* and *only*.
- VARIETY: The varieties studied: *HKE* versus *IndE* versus *PhiE*.
- GENDER: The genders of the speakers: *female* versus *male*.
- AGE: The age groups of the speakers: 14–25 versus 26–35 versus 36–50 versus >50.
- SPKRS: The number of speakers in each group defined by PARTICLE, VARIETY, GENDER, and AGE: between 0 and 170.
- TOKENS: The number of particle tokens found for each group defined by PARTICLE, VARIETY, GENDER, and AGE: between 0 and 74.
- WORDS: The number of words found for each group defined by PARTICLE, 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 $4 \times 2 \times 3 \times 2 = 48$ combinations of all levels of AGE, GENDER, VARIETY, and PARTICLE: 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).

For the re-analyses later, I will not use their values of TMWpaper but the exact ones computed from TOKENS and WORDS, which I will add as a column called just TMW. Also, I am preparing the data a little for all analyses that follow. First, we make *IndE* the reference level of the level predictor VARIETY so that the summary output of any model compares *IndE* separately against the levels of *HKE* and *PhiE*. Second, we change the order of the levels of AGE to an ascending order. Finally, P&F’s approach was to compute “[t]wo regression models [...] in R with relative frequency of clause-final ‘only’ and ‘also’, respectively, as dependent variables, and [AGE, GENDER and VARIETY] as independent variables” (p. 292), so we split up their data frame, which we might call d, into a list, which we might call dd, with two components, each of which contains the data for one particle. On each of the, now, two data sets, P&F performed the following kind of analysis:[2]

> Model selection was conducted using the step function, with *F*-tests as the selection criterion, and allowed for interactions of up to three variables. After model selection, post-hoc Tukey tests (corrected for multiple comparisons) were conducted with the lsmeans function from the eponymous R package. (Lenth & Hervé, 2015)[3]

Thus, P&F’s analyses seem to be summarizable with the following “proxy code”:

TABLE 1 P&F's data as per their appendix 1.

CASE	PARTICLE	VARIETY	GENDER	AGE	SPKRS	TOKENS	WORDS	TMWpaper
C001	also	HKE	female	14–25	84	45	107,796	417
C002	also	HKE	female	26–35	3	0	5304	0
C003	also	HKE	female	36–50	6	0	5888	0
C004	also	HKE	female	>50	0	0	0	0
C005	also	HKE	male	14–25	15	5	16,599	301
C006	also	HKE	male	26–35	2	0	2892	0
C007	also	HKE	male	36–50	4	0	3204	0
C008	also	HKE	male	>50	1	0	902	0
C009	also	IndE	female	14–25	52	74	43,410	1705
C010	also	IndE	female	26–35	33	33	29,456	1120
C011	also	IndE	female	36–50	27	40	25,282	1582
C012	also	IndE	female	>50	9	11	6757	1628
C013	also	IndE	male	14–25	18	14	14,236	983
C014	also	IndE	male	26–35	26	24	25,082	957
C015	also	IndE	male	36–50	37	31	34,771	892
C016	also	IndE	male	>50	37	35	31,927	1096
C017	also	PhiE	female	14–25	170	37	93,197	397
C018	also	PhiE	female	26–35	78	9	25,083	359
C019	also	PhiE	female	36–50	54	6	11,674	514
C020	also	PhiE	female	>50	**0**	**2**	**3613**	**554**
C021	also	PhiE	male	14–25	69	6	32,386	185
C022	also	PhiE	male	26–35	**75**	5	17,283	289
C023	also	PhiE	male	36–50	106	6	6190	969
C024	also	PhiE	male	>50	1	0	3937	0
C025	only	HKE	female	14–25	84	29	107,796	269
C026	only	HKE	female	26–35	3	2	5304	377
C027	only	HKE	female	36–50	6	0	5888	0
C028	only	HKE	female	>50	0	0	0	0
C029	only	HKE	male	14–25	15	1	16,599	60
C030	only	HKE	male	26–35	2	0	2892	0
C031	only	HKE	male	36–50	4	0	3204	0
C032	only	HKE	male	>50	1	0	902	0
C033	only	IndE	female	14–25	52	41	43,410	944
C034	only	IndE	female	26–35	33	20	29,456	679
C035	only	IndE	female	36–50	27	17	25,282	672
C036	only	IndE	female	>50	9	0	6757	0
C037	only	IndE	male	14–25	18	17	14,236	1194
C038	only	IndE	male	26–35	26	14	25,082	558

(Continues)

TABLE 1 (Continued)

CASE	PARTICLE	VARIETY	GENDER	AGE	SPKRS	TOKENS	WORDS	TMWpaper
C039	only	IndE	male	36–50	37	10	34,771	288
C040	only	IndE	male	>50	37	7	31,927	219
C041	only	PhiE	female	14–25	170	12	93,197	129
C042	only	PhiE	female	26–35	78	1	25,083	40
C043	only	PhiE	female	36–50	54	1	11,674	86
C044	only	PhiE	female	>50	0	0	3613	0
C045	only	PhiE	male	14–25	69	0	32,386	0
C046	only	PhiE	male	26–35	**78**	1	17,283	58
C047	only	PhiE	male	36–50	106	0	6190	0
C048	only	PhiE	male	>50	1	1	3937	254

```
analysis.for.also <- step(lm(TMW ~ AGE*GENDER*VARIETY, data = dd$also))
analysis.for.only <- step(lm(TMW ~ AGE*GENDER*VARIETY, data = dd$only))
```

I am using “seem to be” and “proxy code” because these lines would actually not work because these models actually have a perfect fit, which is why the step function will return an error (“AIC is infinity for this model, so ‘step’ cannot proceed”). However, if we use the function MASS::stepAIC with bidirectional model selection and a null model as the starting model as below, we arrive at the same final models as P&F.

They then interpret the results on the basis of post hoc tests for each model (laudably correcting for multiple tests) and visualizations of:

- the 24 observed means of TMW for all combinations of VARIETY, AGE, and GENDER for each particle (the right panels of their figs. 1 and 2);
- the 12 observed differences of female minus male frequency means for all combinations of VARIETY and AGE for each particle (the left panels of their figs. 1 and 2).

They also provide a tabular overview of the results for both particles, which I am not going to discuss here, see Gries (2021) for discussion. Let us now turn to two sets of problems with the hypotheses and their analyses.

2 | PROBLEMS OF THE HYPOTHESES/OPERATIONALIZATIONS

2.1 | Lack of precision

The first set of problems is conceptual, not statistical, but with a very close connection to the regression modeling part below. P&F have two regression models—one for *also*, one for *only*—and each of them can give rise to many different mean TMW-values and their comparisons. For each *also* and *only* separately, there could be:

- a main effect of AGE with 4 means, which permits six post hoc comparisons;
- a main effect of GENDER with 2 means;
- a main effect of VARIETY with 3 means (and perhaps three post hoc comparisons);
- an interaction of AGE:GENDER with 8 means (and perhaps 28 post hoc comparisons);

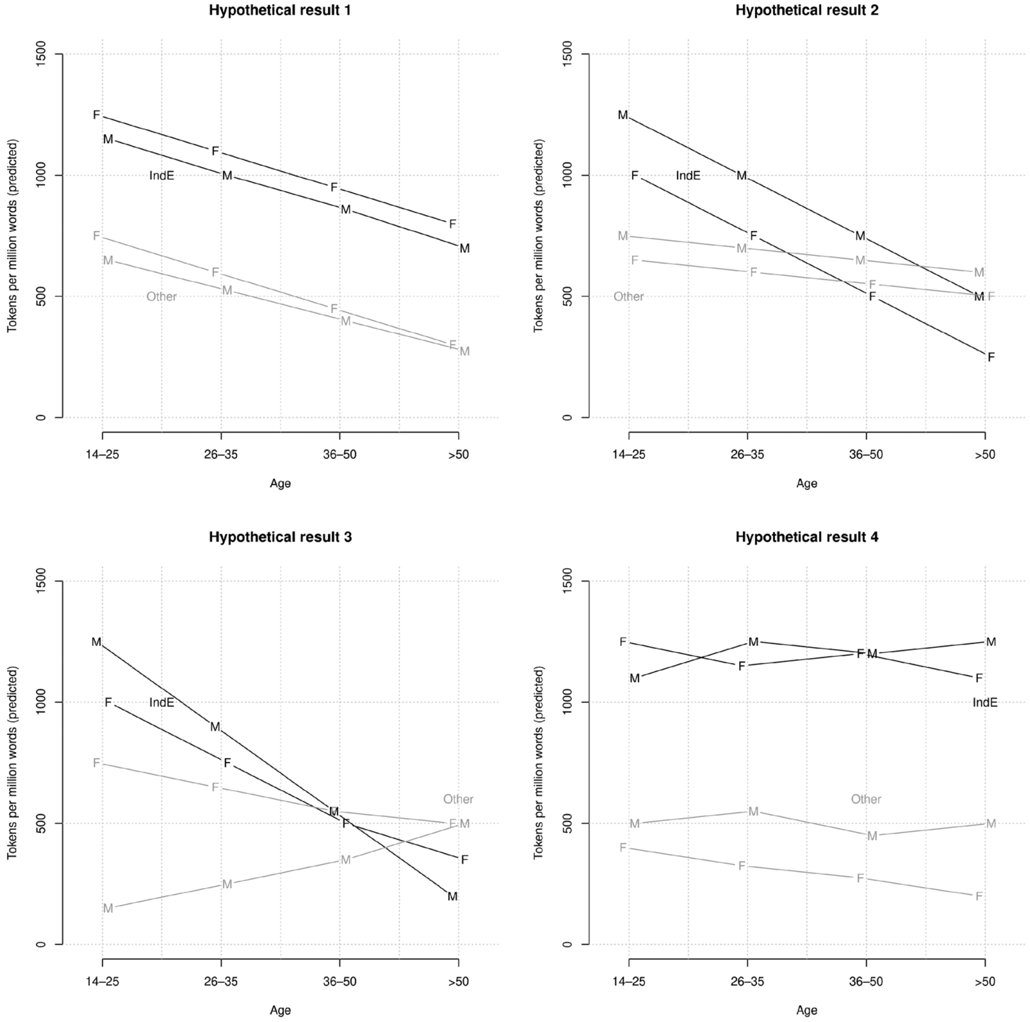

FIGURE 1 Four hypothetical results for AGE:GENDER:VARIETY.

- an interaction of AGE:VARIETY with 12 means (and perhaps 66 post hoc comparisons);
- an interaction of GENDER:VARIETY with 6 means (and perhaps 15 post hoc comparisons);
- an interaction of AGE:GENDER:VARIETY with 24 means (and perhaps 276 post hoc comparisons).

In other words, there is a bewildering number of means and possible comparisons between them but P&F do not formulate very well which means they will compute, which comparisons they will make, and what the results need to be for their hypotheses to be considered "verified". For instance, their formulation "[i]f clause-final 'also' and 'only' are used more frequently by younger speakers than by older speakers" sounds like a directional prediction that the regression model for each particle will feature a main effect of AGE as an ordinal trend. They also say "and/or more frequently by female speakers than by male speakers, […]," which sounds like a directional prediction that the regression model for each particle will feature a main effect of GENDER. And interpreting their language that way seems to be supported by their later phrasing, for example, "we expected to find that younger speakers use these [clause-final particles] more often than older speakers [main effect of AGE], and that female speakers use them more often than male speakers [main effect of GENDER]". But what does "and/or" mean? Is one of the two main effects enough for them to consider their increased-frequency hypothesis to be confirmed? Do we need both? What if the two results are in opposite directions? Or does "and/or" mean the two predictors should interact? May interact?

In fact, the situation with regard to a perhaps expected main effect of GENDER is even trickier because on p. 290f., they discuss the role of GENDER for language change. Specifically, they state that women are often faster than men to adopt linguistic changes (which means we might expect more clause-final particles from women in their data). However, they also state that "this often applies to features that have overt prestige, whereas men are more likely to favour features with covert prestige" but that "[a]t present, there are no attitudinal studies that have examined whether covert or overt prestige is assigned to the use of clause-final 'also' and 'only' by speakers of IndE". In other words, as P&F do not commit to a level of prestige to attach to clause final *also* and *only*, we actually do not know what to expect regarding their use by speakers of different genders, which also means we do not know whether this discussion of prestige walks back the expected gender effect or not.

Then, they also say "if the relative proportion of the use of innovative 'also' and 'only' by young (and) female speakers is higher than that of older (and) male speakers in both PhiE and HKE (when compared with IndE), this would indicate that the use of the feature is a more recent innovation in the former two varieties". This statement is completely different from the previous one because this one predicts an interaction in each particle's model, namely the potential interaction of GENDER "(and/or)" AGE and VARIETY. However, the parenthesized *and/or* in their formulation does again not commit to a clear expectation: Are they predicting that young female speakers will behave differently from old male speakers? From all old speakers (regardless of gender)? From all male speakers (regardless of age)? That young speakers of both genders will behave differently from old speakers of both genders? And how does any of these relate to this: "in both PhiE and HKE (when compared with IndE)"? This can be read as two comparisons (comparing IndE first to HKE and then to PhiE) or as one (comparing IndE to the combination of HKE and PhiE, which would create even more means than those already listed above).

Finally, and to set the stage for Subsection 3.2.2, they are also ambiguous with regard to whether they expect the particles to behave differently. On p. 291, they say "if the relative proportion of the use of innovative 'also' and 'only' [...]," making no distinction between the two particles, but on p. 290 they discuss innovative *only* and continue with "The case of innovative 'also,' however, is more complex," which implies a difference between the particles. Again, the reader wonders what exactly the expectations for the model(s) are: are the particles supposed to behave identically or not? And supposed to behave identically with regard to what (other variables)?

I know how pedantic the above sounds but consider Figures 1 and, which represents four possible outcomes of how the predictors AGE, GENDER (*F* vs. *M*), and a binary version of VARIETY (*IndE* vs. *other*) might behave (i.e., I am averaging across particles). Now, from all the quotes regarding P&F's two sets of expectations—the *sociolinguistic expectations* regarding AGE and GENDER and the *epicentral expectation* regarding IndE—can we decide (i) for each hypothetical result whether it would confirm either of their sets of expectations and (ii) which of the four hypothetical results would do so most convincingly?

I certainly am not able to. Are we supposed to compare slopes of AGE lines? For each GENDER separately or both combined? Do they both need to go down or not? Do the lines for IndE need to remain above those for the others (see next section)?, and so on. Whichever way one looks at it, the language of the hypotheses is so unclear that, from their more theoretical/linguistic parts, it is impossible to know how they will test their hypotheses in the regression models that follow and how which results from the regression models will make them decide what the outcomes mean. We do not know what they require to consider their expectations confirmed: a main effect of AGE, a main effect of GENDER, or both, an interaction such as AGE:VARIETY or GENDER:VARIETY or AGE:GENDER:VARIETY? The fundamental idea of null hypothesis significance testing is to formulate alternative and null hypotheses that cover all theoretically possible results so that, once one has actual results, one can use them to decide to adopt which hypotheses, but here just about every expectation is formulated so vaguely, that it never becomes clear how any given result will actually be interpreted.

2.2 | What do between-variety frequencies reveal?

Another assumption in P&F is that the frequencies of clause-final particles in the presumed epicentral variety (IndE) are (significantly) different from the frequencies in the other varieties (HKE and PhiE). But what can really be inferred from such frequency differences? As Bernaisch et al. (2022, section 2.1.2) discuss in more detail, there is no good reason to assume that a feature X might spread from an epicentral variety V_{EC} to others but will always be more frequent in V_{EC}; in fact it is not clear one can assume *any* specific relation between such frequencies in varieties. If feature X has a frequency f_X in an epicentral variety V_{EC}, that does not predict anything:

- because of the influence of V_{EC}, f_X might grow in other varieties, but
  - f_X in the other varieties might never overtake the frequency in V_{EC};
  - f_X in the other varieties might clearly overtake the frequency in V_{EC};
- in spite of V_{EC} being a clear epicenter, f_X might not change in other varieties.

Thus, even if IndE was an epicenter, the frequencies of clause-final particles might still remain low in HKE and PhiE (for reasons not captured in, i.e., variables not included in, the current regression models). Same thing the other way round: what frequencies of clause-final particles would P&F need to find to support their claim that their data support IndE as an epicenter? Do clause-final particles need to be less frequent in HKE or PhiE than in IndE? The whole apparent time or just part of it? More frequent? Or are both ok as long as the frequencies in HKE and PhiE are still growing? What if one or both had already leveled off? And, as we will discuss in more detail below, the use of these amalgamated frequencies is theoretically uninterpretable anyway because it neither includes any linguistic or structural features nor any information about individual variation. Thus, while I do not know what to expect myself for clause-final *also* and *only*, proper regression modeling requires the modelers to formulate predictions that are clear and that are clearly (dis)confirmable by sizes and signs of significant regression coefficients such that the researcher and their audience can say "if mean A is greater than mean B or slope M is greater than slope N, then hypothesis_1 is supported—otherwise, it is not"; this is something we do not find here.

3 | PROBLEMS OF THE STATISTICAL MODELING

3.1 | A few minor mistakes

Before we turn to the most important modeling issues of this paper, let us briefly mention a few issues with the analyses that, while less useful from a bigger-picture didactic perspective, are still noteworthy if only to make sure they do not get emulated in future work by others.

3.1.1 | Errors in the data

It seems as if the appendix providing the data contains some errors and it is unclear whether their analyses were performed on the data as shown or the correct version. For instance, P&F report two different numbers for what should be one and the same number of speakers, namely cases 22 and 46 (the number of male PhiE speakers between the ages of 26 and 35 producing 17,283 words), see Table 1.

Also, how can a certain configuration of VARIETY, GENDER, and AGE not be manifested by a single speaker, but still yield 3613 words and two clause-final *also*s (case 20)?

3.1.2 | Under-reporting and under-analyzing

There are a few ways in which P&F under-report/under-analyze the data. As for under-reporting, readers do not get to see the customary summary tables of their final models for either particle: none of the information that is customarily provided—coefficients, standard errors, *t*-values, confidence intervals—is offered here. Also, we get no goodness-of-fit results: what are the R^2s for m.final.also and m.final.only? Additionally, their description does not make it completely obvious what post hoc tests they computed—only replicating the analysis oneself shows that they did something reasonable, namely, for example, for *also*, testing (1) the three differences between the levels of VARIETY within each GENDER and (2) the one difference between the levels of GENDER within each VARIETY—in other words, it is good that they did not exhaustively tested all 15 differences between all levels of GENDER crossed with all levels of VARIETY. Finally, I personally would have preferred to have some visual aids representing the results better than several very dense paragraphs full of individual post hoc tests (because the graphs they provide are far from ideal, see the next section).

As for under-analyzing, P&F treat the predictor AGE as categorical, ignoring its ordinal information, something that I frequently see in manuscripts that I review: The levels of AGE are not just categorically different, they can be ranked, and P&F's hypotheses regarding AGE should actually have made them very interested in ordinal effects.

3.1.3 | Over-reporting

At the same time, the analyses also involve some over-reporting. For example, P&F's analysis for *also* returns a final model with the interaction of VARIETY:GENDER as the highest-level predictor. Correspondingly, the interpretation of m.final.also should center on the frequency means of the six different combinations of three varieties and two genders, but instead we are offered all 24 observed means and an additional representation of the 12 differences between female and male means (and all of those without confidence intervals)—why are we given 24 numbers (and 12 more) rather than 6, if the model tells us that most of those 24 are not relevant? Same for the analysis of *only*: The final model has the interaction of VARIETY:AGE as the highest level predictor and no effects of GENDER, which means we should see a discussion of 12 means (three varieties times four age groups), but we again are given 24 observed means and, for a model about which the authors say GENDER is irrelevant, 12 gender-specific differences! Providing many times more means/differences than are actually supported in models only obfuscates the results; a focus on only the predictors that actually seem to have an effect—after all, finding those was the purpose of the modeling—would have been more useful to readers.

3.2 | Four fundamental issues

Let us now turn to the four most fundamental issues that have a much bigger impact on P&F's analysis in particular and pervade many other analyses in general. In each section, I will first provide a brief theoretical introduction to the core of the problem, then I will explain how it is manifested in P&F's analyses.

3.2.1 | Issue 1: No level-1 predictors

General introduction

In many regression modeling contexts involving learner corpus research or varieties research, we can distinguish predictors in terms of the level on which they are observed and entered into the analysis. Imagine a corpus-based study of a structural alternation such as the genitive alternation (*of* vs. *s*) or the dative alternation (ditransitive vs. prepositional datives). In such studies, the response variable is usually each linguistic choice of one of the alternants in the analyst's sample. The level at which that response variable is studied can be referred as "level-1," and predictors measured at that level, therefore, pertain differently to each individual choice and can therefore be called **level-1 variables**. For instance, we know that the genitive alternation is affected by POSSANIM, that is, whether the possessor is animate or not, and for every genitive choice—*of* or *s*—we can annotate the "possessor" for this level-1 predictor. Similarly, we know that the dative alternation is affected by the relation of the patient's and the recipient's length and for every choice of a ditransitive or prepositional dative we can identify the relevant lengths and compute the difference of their (logged) lengths to arrive at a level-1 predictor, which might be called LENGTHDIFF. Level-1 predictors are the contextual, linguistic, structural, psycholinguistic predictors one would typically think of if one is asked "what do you think are the factors affecting the constituent order alternation X?": length, animacy, priming, givenness/topicality, definiteness, specificity, …

However, in many research contexts, we also have higher-level variables. For example, more and more studies recognize the importance of including in an analysis which speaker or stimulus a certain linguistic choice "belongs to". For example, in a corpus-based study of the genitive alternation, one learner might contribute six data points, for example, four *of*-genitives and two *s*-genitives. Each of these comes with values on the level-1 variables, but they are all nested into the **level-2 variable** SPEAKER: If you know the specific corpus example, which is attested in one and only one file, you know the one speaker who produced it. And in cases where each speaker contributes just one text to the corpus (but each text could still contain multiple instances of one or both genitives, this level-2 variable could be called TEXT. Such level-2 variables must be included (often as random effects) for both conceptual reasons (e.g., the importance of individual variation for a phenomenon) and for statistical reasons (because the data points of one speaker/text have more in common with each other than with those of other speakers; in a sense, they instantiate repeated measurements).

In addition to level-2 variables, we can also have **level-3 variables**. In learner research contexts, the learners' L1s could be a case in point: In a scenario where each speaker has one L1 and contributes one text, which may contain multiple genitives, we could represent this as follows: GENITIVE$_{\text{level 1}}<_{\text{nested into}}$ SPEAKER/TEXT$_{\text{level 2}}<_{\text{nested into}}$ L1$_{\text{level 3}}$. Thus, SPEAKER would be nested into L1 because, if you know the speaker, you know their L1. Same for varieties research: the speaker's variety is a level-3 variable because the level-2 variable SPEAKER is nested into it: if you know the speaker, you know the one variety they belong to. However, it might also be the case that a speaker contributes multiple texts, in which case we would face this situation: GENITIVE$_{\text{level 1}}<_{\text{nested into}}$ TEXT$_{\text{level 2}}<_{\text{nested into}}$ SPEAKER$_{\text{level 3}}<_{\text{nested into}}$ L1$_{\text{level 4}}$, and so on. A frequent example of such multi-level modeling studies is how in educational studies:

- behavioral responses such as responses in a 10-question test taken by students are the response variable at level-1;
- there, STUDENT would be a level-2 variable recorded to account for individual variation; similarly, variables such as BOOKSATHOME or HOURSSELFSTUDY would also be level-2 variables because they pertain to—that is, describe—the student and all their responses rather than a specific test response;
- CLASSROOM would be a level-3 variable recorded to account for variation between classrooms; similarly, if all classrooms had different teachers, then TEACHER would be a level-3 variable;
- SCHOOL would be a level-4 variable recorded to account for, say, how well a school is funded by its district;
- SCHOOLDISTRICT would be a level-5 variable recorded to account for, say, socio-economic differences between school districts, and so on.

This is very important, because, on a general level, this discussion should make one recognize that much of the empirical work in 20–30 years of corpus research has restricted itself to "studying" highly local and highly context-dependent level-1 linguistic phenomena—choices to produce or not produce a certain word, grammatical marker, and so on, or choices to produce one of two or more (syntactic) alternatives in some context—as if those were really meaningfully explainable with reference to level-2 and/or level-3 variables alone. Does anyone really think the individual responses in a test administered to students would be explained well by just aggregating the data on the level of the school or the school district? Would anyone really think that the genitive alternation would be explained well by just aggregating constructional frequencies on the level of the dialect of the speakers? If the answers to these questions are *no*, then why do many corpus studies study the frequency of a certain linguistic choice (a level-1 phenomenon) in data from speakers of different L1s or varieties using only a level-3 or -4 predictor such as L1/VARIETY?

The logic underlying the answer to that question is actually really straightforward even though it is hardly ever made explicit and also wrong. It is explicated here for learner corpus research, but the same holds for varieties research:

- learner corpus researchers are interested in "comparing/contrasting what non-native and native speakers of a language do in a comparable situation" (Péry-Woodley, 1990, p. 143, cited by Granger, 1996, p. 43);
- the essays written by the learners from various L1s and the essays written by the native speaker students were written in similar language-production settings (e.g., timed essay-writing situations given a certain prompt);
- this "similarity of language-production setting" permits us to just assume that "all other things are equal" and we can therefore meaningfully compare and interpret the frequencies with which certain linguistic decisions are made by native speakers (NS) and different kinds of learners (NNS).

As Gries and Deshors (2014, p. 113) have argued, however,

> [i]t is easy to see that this seems quite unrealistic: for example, the choice of the modal verbs *can* versus *may* is determined by fifteen or so different factors, $F_{1\text{-}15}$, including the syntactic characteristics of the clause and various morphological and semantic features of the subject […], and perhaps also by the circumstances of production, which we may call 'register'. Thus, the traditional interpretation of 'in a comparable situation' leads to the somewhat absurd assumption that we compare uses of NS and NNS that are completely different in terms of $F_{1\text{-}15}$ and only share the single factor that they were produced in an essay-writing situation in school.

This brings us to P&F.

Application to P&F

By now it should be clear how this relates to P&F (and in fact many other variety studies such as Yeung, 2009; Davydova et al., 2011; Bruckmaier, 2017, to name a few examples): They are trying to explain the frequency of level-1 choices (whether or not to put *also* and *only* clause-finally) on the basis of two level-2 variables pertaining to speakers (AGE and GENDER) and one level-3 variable (VARIETY), but ignore any and all linguistic/contextual level-1 predictors as well as individual variation. In other words, they are conflating all level-1 decisions of a group of up to 170 speakers defined by level-2 variables AGE, GENDER and the level-3 variable and VARIETY into just a single proportion. This amounts to a *massive* loss of information due to ignoring everything that linguists usually care about, namely contextual, linguistic, structural, and psycholinguistic predictors, and individual speakers.[4] Imagine that, for processing-related reasons, clause-final focus particles are more likely in negated clauses. Imagine, further, that IndE exhibits twice as many clause-final focus particles than HKE and PhiE. What can one make of that latter finding? Nothing, really, because the higher mean for the level-3 variable VARIETY: *IndE* does not correct for whatever effect the level-1 predictor NEGATION has in such observational data:

- if negation was, say, 1.9 times as frequent in the IndE data than in the HKE/PhiE data, then most of the twice-higher frequency of clause-final focus particles in IndE is readily explainable by the similar proportion of negation—VARIETY as a predictor would not be needed;

- if negation was, say, *less* frequent in the IndE data than in the HKE/PhiE data, then NEGATION cannot be the driving force for the high frequency of clause-final focus particles in IndE—instead, it could be VARIETY (and/or of course other factors).

Thus, in the absence of (proper control of) level-1 predictors, higher level effects can *never* be taken at face value because they might be entirely due to the ignored level-1 effects. Thus, given that P&F's analyses do not control for linguistic/contextual factors on level 1, they can *by definition* not be certain that any patterns in the data they are happy to ascribe to their level-2 predictors AGE and GENDER or their level-3 predictor VARIETY are really due to those, and not even the many additional statistical improvements suggested below can avoid this inconvenient truth. The only way to make the kind of inferences they are interested in is to include level-1 predictors and to account for the three-level structure of the data (and Gries & Adelman, 2014, but especially Gries & Deshors, 2015, discuss an alternative statistical way to do so that has been adopted in a variety of [learner and variety] studies).

3.2.2 | Issue 2: Two models on one data set

General introduction

If one studies some phenomenon with a multifactorial regression model, one uses multiple predictors (variables in whose (often assumed causal) effects one is interested in) and perhaps control variables (variables included to control for their potential effect even if they are not what the study is focused on). That means one needs to formulate a regression model that embodies one's hypotheses as well as possible. Imagine you did a corpus-based study of verb-particle constructions (VPCs) (*He picked*$_{\text{V}}$ *up*$_{\text{PART}}$ *[the book]*$_{\text{DO}}$ versus *He picked*$_{\text{V}}$ *[the book]*$_{\text{DO}}$ *up*$_{\text{PART}}$) and had these predictors:

- LENGTHDO: A level-1 predictor representing the length of the direct object in words.
- LITERAL: A level-1 predictor representing whether the meaning of the VPC is perfectly literal (e.g., *He threw up the ball to the ceiling*) or not so much (e.g., *He threw up his lunch*).
- DIRECTIONALPP: A level-1 predictor representing whether the VPC is modified by a following directional PP (e.g., *He picked the ball up [*$_{\text{PP}}$ *from the ground])* or not (e.g., *He picked the ball up and left*).
- DIALECT: A level-3 predictor representing whether the VPC was from American English (AmE) or British English (BrE).

As soon as one has multiple predictors/controls, one has to decide how these behave together: are their effects additive or do variables interact with each other? Perfectly **additive behavior** would mean that each variable's effect is independent of every other variable's effects. Examples of additive behavior could be that:

- literal VPCs make V-DO-PART 25% more likely than V-PART-DO *regardless* of the length of the DO, *regardless* of whether there is a following directional PP or not, and *regardless* of the dialect;
- following directional PPs make V-DO-PART 35% more likely than V-PART-DO *regardless* of the length of the DO, *regardless* of whether the phrase is literal or not, and *regardless* of the dialect;
- and if these two predictor levels "come together," then they make it 25+35 = 60% more likely to have V-DO-PART than V-PART-DO—the effects just add up.

By contrast, **interactions** mean that one variable's effect is dependent on what other variables' effects are. An example of interactions could be that:

- literal constructions make V-DO-PART 25% more likely than V-PART-DO;
- following directional PPs make V-DO-PART 35% more likely than V-PART-DO;

- but if these two predictor levels come together, then they make V-DO-PART 95% more likely than V-PART-DO—the effects do not just add up to 60%, their confluence *amplifies* the tendency to have V-DO-PART.

The opposite example would be:

- literal constructions make V-DO-PART 25% more likely than V-PART-DO, but only when the DO is shorter than 5 words;
- literal constructions make V-DO-PART 5% more likely than V-PART-DO, when the DO is 5+ words long;
- meaning, here LENGTHDO *weakens* the effect of LITERAL.

How would we study this statistically? For instance, how would we determine the role of DIALECT in our example? Two approaches might come to mind and, unfortunately, one of them is frequent yet frequently wrong. The first approach would be to fit two models: one to the American data with, say, LENGTHDO, LITERAL, and DIRECTIONALPP as predictors, the other one is then the same model fit to the British data.

```
model.AmE <- glm(CONSTRUCTION ~ LENGTHDO + LITERAL + DIRECTIONALPP,
data = DIALECT = = “AmE”, …)
```

```
model.BrE <- glm(CONSTRUCTION ~ LENGTHDO + LITERAL + DIRECTIONALPP,
data = DIALECT = = “BrE”, …)
```

The second approach would be to fit one model to both dialects, but include DIALECT as a predictor in such a way that we can determine whether it affects/moderates—amplifies or weakens—the effects of LENGTHDO, LITERAL, and DIRECTIONALPP because we include DIALECT’s interactions with these level-1 predictors.

```
model.bothE <- glm(CONSTRUCTION ~ DIALECT * (LENGTHDO + LITERAL + DIRECTIONALPP), …)
```

As discussed in Gries (2021, esp. sections 5.2.4 and 5.2.8, and the exercises for chapter 5), the former approach is, while still used too often, incorrect. If one fits such two models (represented here schematically), then, while the numerical results for each predictor in each separate model will surely be different, one cannot see straightforwardly whether they are significantly different from each other or not. This is for the obvious reason that models can only compare what they know of (and are instructed to compare), and while model.AmE will return a coefficient for LENGTHDO, “it doesn’t know about the results for when DIALECT = = ‘BrE’” and can therefore not compare the two, which means an analyst cannot straightforwardly see whether the effect of LENGTHDO is “the same” in AmE and BrE or not.

The second approach deals with this much better: If the effect of LENGTHDO is “the same” in AmE and BrE, the interaction DIALECT:LENGTHDO will be not significant, if the effect of LENGTHDO is different in AmE and BrE, the interaction DIALECT:LENGTHDO will be significant, and this will be straightforwardly visible from the summary output of the regression model.

Application to P&F

It should again be clear where this is headed. To the best of my abilities, I was not able to find any justification for P&F conducting separate analyses for the particles, but I doubt any such justification would have been acceptable anyway: Although their hypotheses are not formulated clearly enough for us to know, if they had expected “clause-finality” to be instantiated *differently* frequently for each particle, they should have fitted one model and included PARTICLE as a predictor to see whether their expectation was borne out by PARTICLE having a significant effect on its own (or in an interaction, see below). And if they had expected “clause-finality” to be instantiated *equally* for each particle, they should still have that same model with PARTICLE as a predictor to see whether their expectation was borne out by PARTICLE having no significant effects and, thus, getting deleted during their model selection process. There simply is no good reason to not fit one model with PARTICLE as a predictor. In fact, the argument must even be extended to include the other predictors. Because, really, PARTICLE should not only be included as an individual predictor or main effect—given their approach, it should be permitted to interact with everything else. More precisely, they allowed every predictor to interact with everything else in the model: Separately for each particle, their model selection process considered the interaction of AGE:GENDER:VARIETY so that they would be able to see whether:

- the effect of the AGE:GENDER interaction is constant across varieties;
- the effect of AGE:VARIETY is constant across genders;
- the effect of GENDER:VARIETY is constant across ages.

Thus, it would have only been consistent to fit the one big model with PARTICLE potentially interacting with everything else:

```
summary(m.final.all <- stepAIC(
   # note the new data argument: all the data from both particles now
   lm(TMW ~ 1, data = d),
   # bidirectional model selection …
   direction = “both”,
   # between these 2 extremes: lower null & the new upper/full model:
   scope = list(lower = ~1, upper = ~AGE*GENDER*VARIETY*PARTICLE),
   trace = 0)) # do not show all steps
```

This way they could have seen whether:

- the effect of the AGE:GENDER:VARIETY interaction is constant across particles, and, if not,
- the effect of AGE:GENDER is constant across particles, and/or
- the effect of AGE:VARIETY is constant across particles, and/or
- the effect of GENDER:VARIETY is constant across particles, and so on.

This would be the natural way to apply the logic of their models to the new single big model, and it would allow them to see whether everything behaves the same for each of the two particles. This would clearly be relevant linguistically, and, as mentioned above, there is no good reason to not do this statistically.

3.2.3 | Issue 3: Not weighting observations

General introduction

The vast majority of regression modeling studies in linguistics or in corpus linguistics targets a response level representing a single linguistic choice: the (level-1) choice of a particular word (over others), the choice of a particular grammatical construction (over others), the choice to lengthen a syllable (over not lengthening it), and so on. Such data are usually represented, and entered into statistical analysis, using the **case-by-variable format**, a format in which, typically,

- each row represents one data point, that is, one level-1 observation capturing a speaker’s decision in a response variable;
- each column represents a variable whose levels describe the level-1 observation.

TABLE 2 Three cases from P&F's data.

CASE	PARTICLE	VARIETY	GENDER	AGE	SPKRS	TOKENS	WORDS	TMWpaper
C017	also	PhiE	female	14–25	**170**	37	93,197	**397**
C018	also	PhiE	female	26–35	**78**	9	25,083	**359**
C026	only	HKE	female	26–35	**3**	2	5304	**377**

TABLE 3 Three cases from P&F's data.

CASE	PARTICLE	VARIETY	GENDER	AGE	SPKRS	TOKENS	WORDS	TMWpaper
C001	also	HKE	female	14–25	**84**	45	107,796	417
C003	also	HKE	female	36–50	**6**	0	5888	0
C005	also	HKE	male	14–25	**15**	5	16,599	301

Returning to the example from above for a moment, if we had 100 VPCs, the case-by-variable format would require, minimally, a data frame with:

- 100 rows, with each one representing one choice of a VPC in our sample, be it a case of *V-PART-DO* or *V-DO-PART*, and its descriptors in the columns;
- 5 columns: one with the response variable and four with the predictors we discussed: LENGTHDO, LITERAL, DIRECTIONALPP, and DIALECT.

This way, the, say, fifth value of CONSTRUCTION would be the fifth constructional choice and the fifth values of LENGTHDO, LITERAL, DIRECTIONALPP, and DIALECT would describe the circumstances of its production. As every row of the data frame would represent the same number of cases/data points—just 1—if we fitted a regression model on this data frame, every data point would have the same impact or weight when it comes to computing the regression model. But, clearly, that is not what is going on in P&F.

Application to P&F

Recall the particle-specific data frames that P&F ran their analyses on; see a few rows from their data again in Table 2.

The range of all TMW-values in the data is ≈1705 (the interval is [0, 1704.7]) while the range of the TMW-values in these three data points is ≈38.2 (a mere 2.2% of the overall range), which can be interpreted as "against the background of all data, these three values are very close together". However, the three fairly similar TMW-values are based on extremely different numbers of speakers: The TMW-values of 397.0085, 358.8088, and 377.0739 summarize the behavior of 170, 78, and 3 speakers, respectively, which also means they probably come with very different degrees of variability. However, P&F's analysis weights all TMW-values the same. To show very intuitively that this is not a good methodological choice, consider the following question: If you read the results of two identically designed polls trying to predict the outcome of the next federal election and the polls are based on 17 and 1000 people (the same ratio as 3-to-170 speakers), would you treat the two polls as equally informative? Or would you give more interpretive weight to the larger poll because, given its size, it is likely to be less volatile? I think the answer is clear—the second of course, that is why standard errors are computed with the sample size in their denominator, for instance. But P&F's analysis does the first because they are conducting the default kind of linear model giving each row the same weight (as if all rows were a single level-1 observation rather than level-2/-3 aggregates of differently many level-1 observations).

One way to analyze the data that could be considered at least slightly better would involve giving each of P&F's rows a weight in the regression analysis that is proportional to, for instance, the number of speakers summarized in that row. One overly explicit way to do this is the following.[5] Consider the three rows of the *also* data shown in Table 3.

One way in which one could make the regression model "see" how many speakers each row represents—84, 6, and 15—consists of creating a new, disaggregated data frame from this one, one that contains the first row of this little three-row sample 84 times, the second row 6 times, and the third row 15 times. That data frame can then serve as the input to the default of an equally weighted linear model to that new data frame. In R, this can be done quickly with maximally three lines of code, and if we then fit P&F's final model for *also*—just to pick one example—we get regression results that are quite different from the results that P&F reported (for *also*). Although this analysis is not unproblematic either,[6] the logic of it should at least raise some awareness of the importance of the different weights with which values can come, which the original analysis did not.

3.2.4 | Issue 4: Potentially wrong regression model

One of the main ways in which regression models are distinguished is based on their response variable. Binary response variables are often analyzed with some form of a binary logistic regression, categorical response variables with 3+ levels are often analyzed with some form of a multinomial regression, ordinal response variables are often analyzed with some form of an ordinal regression. For numeric response variables, there are multiple options, and linear models, the type P&F fit, are among the most widely used ones (e.g., for reaction times, word durations, formant frequencies, etc.) However, the numeric response variable in P&F's data is a (normalized) frequency, meaning it cannot become negative: like an odds value, it is bound to fall into the interval $[0, +\infty]$. Such response variables *may* in certain circumstances be analyzed with the frequent default of a linear model, but if one does that, one needs to show (e.g., with regression diagnostics) that this is actually justifiable for one's data and that the more usual kinds of models for frequency data—for example, Poisson regression or negative binomial regression—were really not required. In this particular case, we are not offered this kind of justification; a cursory exploration suggests that, ignoring all of the issues discussed so far, a linear model here may not be too problematic: the residuals of each model do not differ massively from normality and return nonsignificant tests for nonconstant variance. Still, one would have had to be shown this in the paper, and once one begins correcting the other issues mentioned above by, for example, merging the two particles into one data set, one does in fact encounter massive nonconstant variance problems, which might ultimate necessitate a different regression modeling approach.

4 | RE-ANALYZING THEIR DATA

4.1 | Methods

Let us put this all together and see what addressing at least some of the above issues does to the results. Crucially, we cannot address everything because P&F's appendix 1 does not provide the data to switch to an analysis of level-1 observations, which (i) would allow us to include level-1 predictors governing the choice to position a focus particle clause-finally and (ii) would permit us to look at speaker-specific results.[7,8] Thus, the actually required kind of analysis—some kind of mixed-effects or multilevel model as shown in the next code block—we cannot do:

```
# this could be fitted on all level-1 uses of also/only, clause-final or not:
m.required <- glmer(
  CLAUSEFINAL ~                          # binary response: no versus yes
  1 +                                    # intercept
  PARTICLE*AGE*SEX*VARIETY +             # predictors & interactions
  # minimally necessary random-effects structure (more is likely needed)
  (1|SPEAKER),                           # speakers
  family = binomial,                     # the response is binary
  data = unaggregated.dataframe) # a new unaggregated data of all also/only
```

For this didactically motivated review paper, we will have to work with what we have (ignoring issue 3). For comparability with P&F, I will not correct the errors in their provided spreadsheet and I will ignore issue 4 and also use a linear model. I will, however, improve on their analyses by

1. improving some predictors by using
   a. contrasts reflecting the ordinal nature of AGE;
   b. planned orthogonal contrasts pitting the suspected epicenter of IndE against the other two varieties combined;
2. improving the model by
   a. fitting one big model …
   b. on the expanded data, and
   c. using PARTICLE as an additional predictor that can interact with all others;
3. changing the model selection process by
   a. not using automatic model selection;
   b. limit the degree of interactivity (mostly for didactic simplicity, I would not include all possible interactions in the model);
4. improving the reporting by
   a. providing at least one kind of index of model fit (R^2s);
   b. generating effects plots of predicted values to interpret the results more easily rather than reporting redundant observed frequencies.[9]

Let us do model selection and fit this as our first model (I am not showing any output, see the code file for that):

```
m.all.expanded <- lm(TMW ~ 1 + PARTICLE * (AGE + GENDER + VARIETY)^2, data = d.expanded)
```

All three-way interactions are significant, so we would actually not remove any predictors, we would be done.

4.2 | Selected results

What do these results show? Our (final) model seems interesting in how it comes with a very good fit but *everything* about that model of course comes with the big caveats that we did not do the kind of actually required mixed-effects/multilevel analysis and lost all within-group variability. With that big hedging reminder, mult. and adj. R^2 are >0.97 and much higher than the corresponding values in P&F's analyses. Space does not permit a detailed analysis of all findings and, given the hedges, we would not want to overinterpret these results anyway, but it is clear that the patterns in these results are much more complex than P&F's original analyses suggest (plus, as discussed, P&F's expectations were not formulated in ways that make for straightforward interpretation of these results in the first place). We find that PARTICLE interacts significantly with every other 2-way interaction in the model, but we also see that they do not clearly support any of the expectations formulated by P&F. Consider, for example, Figure 2 for the interaction PARTICLE:AGE:GENDER:

- the left and right panel represent the results for *also* and *only*, respectively;
- in each panel,
  - the *x*-axis represents AGE
  - the *y*-axis represents the predicted TMW-values;
  - the (gray) *F*s and (black) *M*s represent the predictions for women and men, respectively (with their artificially small 95% confidence intervals);
  - the physical size of the letters represents the number of speakers for a certain condition.

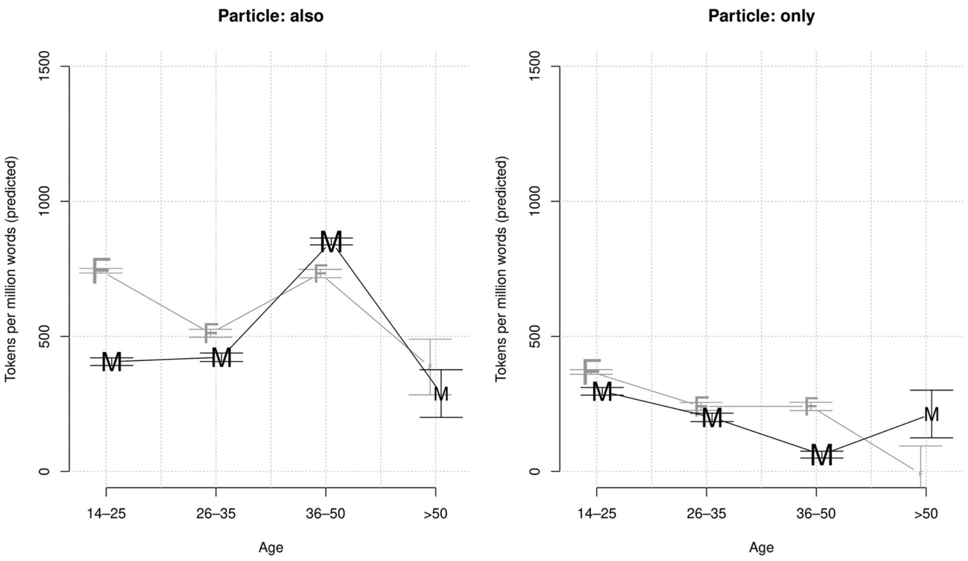

FIGURE 2 The interaction PARTICLE:GENDER:AGE.

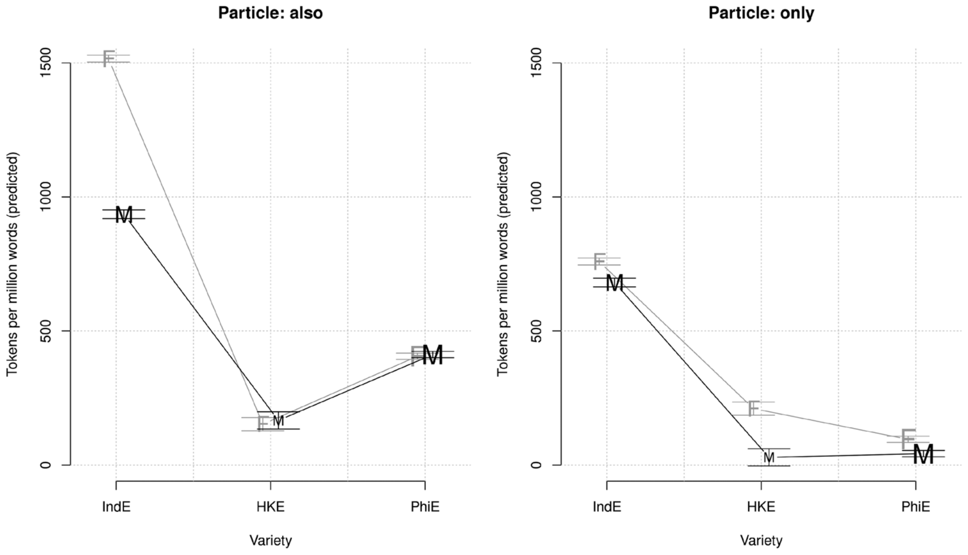

FIGURE 3 The interaction PARTICLE:GENDER:VARIETY.

We can see that there is no real nice ordinal trend of AGE for either gender with *also*, there is a bit of an age trend for women with *only* but not so much for men. The only way to "salvage" an AGE effect is to really cherry-pick results and say that the youngest group always uses clause-final particles more than the oldest, but that is really all there is. As for GENDER, there is a bit of the expected effect: In six out of eight conditions, women use clause-final particles more than men. However, the results are clearly not homogeneous and it is not clear if P&F would consider them as supporting their expectations or not.

Let us also very briefly look at a three-way interaction involving VARIETY in Figure 3.

There is a fairly clear effect of VARIETY such that IndE has most clause-final particles across both particles and genders. The effect of GENDER is trickier, though, because in three to four of six cells, women use clause-final particles more than men, but in the others (*also* in HKE and PhiE, maybe *only* in PhiE) the differences seem practically irrelevant. Additionally, for *also* the combined frequency in HKE is much lower than in PhiE, whereas we seem to find the opposite for *only*. With regard to particles, by contrast, the frequencies of both are very similar in HKE, but very different in IndE (also across genders) and PhiE, where the genders are very similar, but different for the particles.

These results are very complex and the interpretation of this shaky model is beyond the scope of this more didactic paper. However, two things should have become clear. First, part of why they are so difficult to interpret is that we did not get clearly operationalized hypotheses about specific means from P&F, so we do not really know which visual/mathematical comparisons to make in Figures 2 and 3, echoing the difficulties discussed in the context of hypothetical results of Figure 1. Second, the results of this model, which fixes many, but not all, issues of the original analysis, indicate that the original analysis probably substantially underestimated the complexity of the results—it definitely did this in terms of individual variation—and, once that additional complexity is included as well as possible given what is available, the results completely defy simple explanations in terms of the level-2 sociolinguistic predictors of AGE and GENDER and the level-3 predictor of VARIETY.

5 | CONCLUDING REMARKS

Space here was limited: There were several things I could not discuss here as much as I wanted to or at all. However, just to remind everyone, I again urge readers to recognize that this paper is not meant as a hit piece—its purpose was to exemplify several not uncommon yet fundamental problems that can plague regression analyses in our corners of the field and make recommendations that, once one thinks about them, should really be largely or even entirely uncontroversial:

- we need very precisely formulated hypotheses so that we know what kind of results plot will confirm or disconfirm our results. In the present case, we would have needed hypotheses for both the sociolinguistically motivated expectations regarding AGE and GENDER and the epicentrally motivated expectations regarding VARIETY, and they would have had to be very specific:
  - which (means of) frequencies need to be higher or lower than which others to adopt one's alternative hypotheses?
  - which differences or ratios of means need to be higher or lower than which others to adopt one's alternative hypotheses?
  - which intercepts or slopes need to be what to adopt one's alternative hypotheses?, and so on;
- we need to make sure the regression models we want to apply are permitted for the data we have and are able to answer the questions we have for the data;
- we need to include variables at the level of the phenomenon (minimally as controls);
- we need to test for differences using interactions in models on complete data sets rather than fit separate models on parts of the data;
- if we work with aggregate data—which we really should not ever do!—we need to be aware of the weightings our models require, but again, we really should be working with nonaggregated data with level-1 observations and variables and should respect the multi-level structure our data have (here, level-2 speakers being nested into level-3 varieties).

Adherence to such recommendations should help the field reach a more mature level of quantitative sophistication, something that, I hope we can agree on at least that, would be something from which we all would benefit.

ACKNOWLEDGMENTS

I would like to thank the reviewers and editors for their comments and feedback, which helped to make the paper clearer and more explicit in several regards; also, I am grateful to Tobias Bernaisch and Benedikt Heller for much discussion and collaboration over the last few years, from which I have learned a lot.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

ORCID

Stefan Th. Gries https://orcid.org/0000-0002-6497-3958

NOTES

[1] Let me briefly clarify why a specific publication and specifically this one was chosen for the present discussion. First, we indeed need a concrete application because if one merely points out those issues abstractly, as I am often advised to do by diplomatic colleagues, then the reaction is all too often a sea of facial expressions all communicating "here he goes again, pointing out these things no one actually does wrong (anymore)"—exemplification on the basis of concrete, published studies is necessary to move the field along to a better understanding of the methods many of us are using frequently and to better application of these methods. Second, this specific study (laudably) provided all the input data that were analyzed in the paper, something that is only slowly becoming more frequent. Third, some of the mistakes it exemplifies are widespread (as I will point out in the relevant sections). Finally and most generally, the use of predictive modeling techniques in the study of World Englishes in particular is slowly becoming more widespread, so it is important to make sure that the quantitative methods that are published and might therefore serve as role models for future work set good examples. Thus and as I will clarify again at the end of the paper, this contribution is not at all meant to be a hit piece—its ultimate purpose is to make sure we use perform our statistical analyses with the same kind of rigor that we would expect in other disciplines (such as public health or medicine).

[2] Code to run all the analyses is available on the author's website at https://www.stgries.info/research/2024_STG_ProperRegrModlg4VarRes_WorldEngl.html.

[3] This description is actually a little problematic for two reasons. First, P&F say they used the function step for model selection, which is potentially problematic in and of itself because, while there is disagreement about many modeling issues, the vast majority of scholars seem to agree that leaving model selection to an automatic process is hardly ever a good idea. However, P&F also do not specify the direction of model selection they are using—forward, backward, or bidirectional?—although this could potentially lead to different results; it is possible they went with step's default of backward model selection. Second, step is an automatic model selection function that uses *AIC*—not *F*!—as a model selection criterion: P&F are misstating the statistic used for model selection, which is more than just a clerical mistake because *AIC* is considerably more lenient in letting predictors stay in the model than a *p*-value of 0.05 from a standard *F*-test would be; thus, an *AIC*-based model selection process usually leads to models with more predictors than an *F*-test based model selection process; see Heinze et al. (2018, p. 435).

[4] Gries (2024) demonstrates what P&F's approach would amount to in a learner corpus example, namely a reduction of information that the corpus data provide by sometimes more than two magnitudes, a practice which, in his learner corpus case study, also leads to completely wrong results.

[5] There is a less cumbersome way to do this (using weights), but it involves a technicality involving *df*-values and subsequent inferential statistics I want to spare the readers.

[6] The big remaining problem is that, even with this fix, the values of the response variable are now identical for all 84, 6, 15, … repeated rows for each speaker group when the real data are never going to be that homogeneous. In other words, even with the weighting correction, P&F's aggregation of course still lost any and all within-group variation, meaning all inferential statistics are still problematic (see also the documentation for the lm function in R). We will proceed with this for now, but this is a huge and fundamental caveat ruling out this kind of analysis.

[7] I am assuming P&F did not retrieve any contextual/linguistic predictors but "only" the aggregate data provided in their appendix 1.

[8] It is an interesting question of whether one could adopt an alternation-based perspective to their data, meaning whether one could conceive of the response variable being something like POSITION: *clause-final* versus *elsewhere* or something like FOCUSSTRATEGY: *clause-final particle* versus *other(s)*. This is because their review makes it clear that one function of *also* and *only* in Asian Englishes is shared with BrE/AmE (and might thus positionally alternate) while one other function is a "new

semantic meaning" (p. 288) and might therefore not alternate. This is therefore an interesting case resembling the semasiological versus onomasiological distinction: One perspective would start from the presence of the word and look at its position and function, the other would start from the speaker's desire to express some kind of focus and then look at how it is expressed. In this case, however P&F do actually not distinguish these two functions in their models anyway so it is unclear whether an alternation-based reconceptualization and, if so which, would help improve their analyses most or whether other features would be needed.

[9] All these things are explained in more detail in chapters 5 and 6 of Gries (2021).

REFERENCES

Bernaisch, T. J., Gries, S. T., & Heller, B. (2022). Theoretical models and statistical modelling of linguistic epicentres. *World Englishes*, *41*, 333–346.

Bruckmaier, E. (2017). *Getting at GET in World Englishes*. de Gruyter.

Davydova, J., Hilbert, M., Pietsch, L., & Siemund, P. (2011). Comparing varieties of English: Problems and perspectives. In P. Siemund (Ed.), *Linguistic universals and language variation* (pp. 291–324). De Gruyter Mouton.

Gries, S. T. (2021). *Statistics for Linguistics with R* (3rd rev. & ext. ed.). De Gruyter.

Gries, S. T. (2024). Against level-3-only analyses in corpus linguistics. *ICAME Journal*, *48*, 1–25.

Gries, S. T., & 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). Springer.

Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*, 109–136.

Gries, S. T., & Deshors, S. C. (2015). EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research*, *1*, 130–159.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*, 431–449.

Lenth, R., & Hervé, M. (2015). lsmeans: Least-Squares Means. R package version 2.17. Retrieved from http://CRAN.R-project.org/package=lsmeans

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

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

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

How to cite this article: Gries, S. T. (2024). On regression modeling in varieties research. *World Englishes*, 1–21. https://doi.org/10.1111/weng.12694