David Tizón-Couto* and David Lorenz

Variables are valuable: making a case for deductive modeling

https://doi.org/10.1515/ling-2019-0050
Received December 30, 2019; accepted June 14, 2021; published online September 2, 2021

Abstract: Following the quantitative turn in linguistics, the field appears to be in a methodological "wild west" state where much is possible and new frontiers are being explored, but there is relatively little guidance in terms of firm rules or conventions. In this article, we focus on the issue of variable selection in regression modeling. It is common to aim for a "minimal adequate model" and eliminate "non-significant" variables by statistical procedures. We advocate an alternative, "deductive modeling" approach that retains a "full" model of variables generated from our research questions and objectives. Comparing the statistical model to a camera, i.e., a tool to produce an image of reality, we contrast the deductive and predictive (minimal) modeling approaches on a dataset from a corpus study. While a minimal adequate model is more parsimonious, its selection procedure is blind to the research aim and may conceal relevant information. Deductive models, by contrast, are grounded in theory, have higher transparency (all relevant variables are reported) and potentially a greater accuracy of the reported effects. They are useful for answering research questions more directly, as they rely explicitly on prior knowledge and hypotheses, and allow for estimation and comparison across datasets.

Keywords: effect estimation; statistical modeling; theory and data; variable selection

1 Linguists and statistical models: cowboys with cameras

Linguistics has been undergoing a methodological paradigm shift towards an increasingly quantitative discipline (Janda 2013; Kortmann this issue; Sampson

***Corresponding author: David Tizón-Couto**, Department of English, French and German, Facultade de Filoloxía e Tradución, Universidade de Vigo, 36310 Vigo, Spain, E-mail: davidtizon@uvigo.es. https://orcid.org/0000-0003-0788-7954
David Lorenz, Institut für Anglistik/Amerikanistik, Universität Rostock, Rostock, Germany, E-mail: david.lorenz2@uni-rostock.de. https://orcid.org/0000-0002-7451-099X

2005, 2013). This "quantitative turn" has done much to improve the empirical robustness of linguistic research, and perhaps to lead researchers to ask more empirical questions to begin with. Turning to observational data and taking a quantitative perspective is a principled decision; however, which methods and tools to use in analyzing data is often a matter of availability, familiarity, and knowledge of how a given technique can be applied. Most researchers working with language data are interested in learning how to use statistical applications. Many linguists may have read one or several user's manuals that are tailored to their needs, such as Baayen (2008), Johnson (2008), Gries (2013), or Levshina (2015), which explain how to apply statistical analyses to language data. Given the increasing range of possibilities, it is an exciting time to reflect on how statistical methods can help explain linguistic data and enrich our analyses. At the end of the day, however, we – as linguists – are experts on language, not statistics: our interest is in the workings of language, not arithmetic. We want to learn about language, not about numbers. Yet, an informed and linguistically meaningful use of statistical methodology does require some understanding of how it works. Like with any household appliance, we may not need to know exactly where the wires run inside the device, but we need to know what happens when we turn this button or pull that lever.

When a researcher first ventures into the hardware store of quantitative methods, they often find the range of tools bewildering and the complexity of each one of them overwhelming. Bewildered and overwhelmed, they will, perhaps, ask for firm guidelines to direct their decision making. Such guidelines, however, are often of rules of thumb. For example in regression modeling, when asking how many variables we can fit to a set of data points, we are given the "15 events per variable" rule, which states that we need at least 15 observations per variable (Baayen 2008: 195; Harrell 2015: 73). Elsewhere, we read about "10–15 observations per coefficient" (Levshina 2015: 144), and for logistic regression "the less frequent response level divided by 20" (Speelman 2014: 530), by 10 (Hosmer et al. 2013: 407–408; Levshina 2015: 257), or even just by 5–9 (Hosmer et al. 2013: 408; cf.; Vittinghof and McCulloch 2006). What we can really learn from such varying suggestions is not the "rule" so much as the understanding that sparse data will lead to greater uncertainty in the modeled effects. Rather than holding on to firm rules and conventions, we should develop an understanding of the underlying statistical rationales for such recommendations, to eventually explain the reasoning behind the steps we have followed to reach our own statistical conclusions.

A seemingly reliable rule is '$p < 0.05$'; this, we have learned, means statistical significance, and what is significant is worth reporting. However, there has been general criticism against using p-values with an arbitrary cut-off point (such as 0.05), which promotes an inappropriate dichotomous thinking in terms of

"significant versus not significant" (cf. Cumming 2012: 27–33; Vasishth and Nicenboim 2016: 353). This can lead to a bias in model-building, in that variables selected on the basis of *p*-values will result in underestimated uncertainty and overestimated significance (cf. Tong 2019: 247–248). Moreover, *p*-values and null hypothesis testing do not provide a proper interpretation of effects: "[t]he *p* value cannot inform us about the magnitude of the effect of X on Y. Similarly, the *p* value cannot help us to choose which variable explains the most" (Figueiredo Filho et al. 2013: 47; cf.; Cumming 2012; McElreath 2016; Thompson 2002).[1] Again, following a rule (here, $p < 0.05$ as the significance threshold) may be less helpful than forming an understanding of what *p* does *not* tell us.

Moreover, an increasing range of statistical methods is being introduced to the field, and while there is a spirit of novelty and exploration, linguistic research finds itself in a kind of "wild west" of quantitative methodology: there are many new prospects for (statistical) analysis, in the rather well-chartered territory of regression modeling and beyond, such as Correspondence Analysis (Glynn 2014; Greenacre 2007) or conditional inference trees (Gries 2020; Hothorn et al. 2006; Levshina 2021); there are also new methodological frontiers such as mixed effects regression (Gries 2015; Zuur et al. 2009), generalized additive models (Winter and Wieling 2016; Wood 2017) or multidimensional scaling (Borg and Groenen 2005). These new opportunities might have brought along a gold rush, an unrealistic hope for incredible riches (valuable new findings, or perhaps just the gold nuggets of statistically significant effects) to be dug up from the new language data terrain. Especially for the less stats-savvy, this also brings up the perceived need to follow the "customary"/standard procedures specified in textbooks (or previous studies) in order to achieve the state-of-the-art statistical analysis suggested by the reviewers of their publications; but the wild west is not a place to rely on customs and standard procedures. We have to be adventurous and cautious at the same time.

In a nutshell, using statistical modeling in our analyses has brought us closer to the quantitative approaches followed in other scientific fields ("to boldly go where the others already are" [Kortmann this issue]), but it does not necessarily bring us closer to more definite answers as regards the reality we study.

---

1 The downsides of *p*-values and null hypothesis testing have led some scholars to proclaim a paradigm shift in the use and interpretation of statistics. This is evinced in book titles such as *Understanding the new statistics* (Cumming 2012) and *Statistical rethinking* (McElreath 2016). One major proposal is to use effect size estimation and confidence intervals as a statistical yardstick. These can be interpreted directly without necessitating the detour of rejecting (or not rejecting) a constructed null hypothesis. Again, this means that we cannot hold on to a firm rule but instead we get a picture of the size of an effect and the range within which it plausibly falls (cf. Cumming and Finch 2005: 174; Thompson 2002: 31).

1.1 Statistical models as images

"All models are wrong, but some are useful" is a common aphorism in statistics, famously formulated by George Box (1979). In his view, models can only provide useful or illuminating "approximations". We might say they are like the pipe in Magritte's work *The Treachery of Images* (1929). It cannot be packed, lighted, or smoked; it is not a real pipe, only its visual representation (see Figure 1). Similarly, a statistical model offers us a concise abstraction, a singular representation or picture of the linguistic phenomenon under study.

Accordingly, we should bear in mind that a statistical model is not the reality; it is just an (abstract) image of the reality that we study. However, this image should highlight relevant aspects to help us understand the phenomenon of

Figure 1: Magritte's *La trahison des images* 'The treachery of images' (1929), Los Angeles County Museum of Art.[2]

2 © René Magritte, VEGAP, Vigo, 2021 and Digital Image Museum Associates/LACMA/Art Resource NY/Scala, Florence.

interest. We might then conceive of a statistical method as the camera that will allow us to take a picture of the reality we are interested in. The model will take the picture for us, but what the picture shows will depend on the camera's settings, as well as on other circumstances that surround the "reality" and might affect the image indirectly (e.g., natural lighting; noise in the data). The photographer's/researcher's task is to find an appropriate configuration of the camera/model, given the purpose and conditions at hand.

1.2 Variable selection: adjusting the exposure level in our cameras

A researcher will have an overall goal regarding their data, such as pursuing a "confirmatory" or "exploratory" line of investigation, and in many contexts it is essential to state the approach taken (cf. Agresti 2002: 212; Vasishth and Nicenboim 2016). We think that it also helps to consider, for each individual variable, why it is part of the investigation. In multifactorial research settings, different objectives may lead us to include a variable in our study:

(i) Exploration: We want to explore the influence of a variable, but have no particular expectation about whether or in what way that influence will show.
(ii) Confirmation: Previous findings or theory suggest an effect for a variable; we want to test whether the effect is present in our dataset.
(iii) Estimation: We expect a certain effect (from previous findings or theory) and want to evaluate its magnitude in our dataset.
(iv) Arbitration: Previous findings or theories suggest different expectations regarding the effect of a variable; we want to test which of them our data supports.

In settings with multiple input (i.e., independent) variables, our line-up will include predictors of different types. As we proceed to statistical modeling, the way in which a procedure treats a variable is blind to the researcher's original objective – the model does not distinguish between cases (i) to (iv). For example, if we follow through with an automatized variable selection procedure, we end up treating every variable on a par and as dispensable if so judged by statistical criteria (which should only apply to objectives (i) and (ii), if at all). This means we put aside whatever expectations we may have had about the variable on theoretical grounds. Yet, when we discuss our results, the expectations and theoretical objectives may be crucial as the basis of our interpretation of the results. We will argue that these expectations and objectives should also be relevant to the

procedure of data analysis. This means that research aims can and should guide every step of data analysis and, if need be, have priority over automatic processes.

When using a regression model to account for linguistic variation, researchers have to adjust a number of settings when they run an analysis (or "take a picture" of the reality/phenomenon in focus). Deciding on the variables that must be included in the model is the first and possibly the most important adjustment required. To continue with the camera metaphor, we can use automatic settings or make manual adjustments. We might then think of variable selection as the "exposure level" of regression models, i.e., how much light do we need in order to get the shot that comes closest to a reasonable and informative image of the reality? Likewise, when fitting a regression model, a sensible recruitment of variables is required, and the question arises which out of a set of candidate variables are to be part of the model in the end. This is "a question with no definite answer" (Baayen 2013: 347). Variable selection thus has been described as a balancing act "between high within-dataset accuracy on the one hand, and high predictive accuracy for new data on the other" (Johnson 2008: 90). Speelman (2014: 529) adds the warning that an overly detailed model runs the risk of "overfitting to the 'noise' in the data", while a reduced one may lead to "misreading or oversimplifying the patterns in the data".

Textbooks on statistics for Linguistics typically suggest – or at least describe – an Occam's razor approach to model selection, namely the idea that among competing hypotheses, the one with the fewest assumptions should be preferred (cf. Upton 2017: 90). This translates into preferring a model with fewer coefficients – to "get as good a fit as possible with a minimum of predictive variables" (Johnson 2008: 90) – and often implies following an automatic elimination procedure in order to reach a "minimal adequate model" that ultimately drops the variables that do not make a significant contribution to explaining the variance in the dependent variable. There are two popular strategies towards model minimalism: a backward elimination procedure starts with a maximal specification of the statistical model, one that includes all predictors. It then works its way backwards to a parsimonious constellation. Forward selection, on the other hand, builds up towards this minimal specification step-by-step, by adding predictors to the focal model until the minimal adequate set is reached. Forward and backward procedures can also be combined (bidirectional model selection). These techniques feature prominently in the quantitative linguist's bookshelf. For instance, Johnson (2008: 90) suggests stepwise forward variable selection; Baayen (2008) mentions sequential ANOVA tests as a forward stepwise procedure (2008: 199), as well as backwards elimination (2008: 205), but does not explain variable selection techniques in detail. Hosmer et al. (2013: 90–93) promote variable selection and minimal adequate models, although not by purely statistical means but by a

process of "purposeful variable selection" that involves seven different steps. Gries (2013) also suggests (stepwise) variable selection on the basis of Occam's razor (2013: 285), although he does warn that "automatic model selection processes can be dangerous: different algorithms can result in very different results" (2013: 292). Levshina (2015) mentions the option of "retain[ing] all theoretically relevant factors in the model" (2015: 149) and is cautious about stepwise selection procedures (2015: 152); similar words of caution come from Agresti (2002: 214), who states that "algorithmic selection procedures are no substitute for careful thought in guiding the formulation of models". However, most current research papers on linguistic variation that use regression modeling seem to implicitly or explicitly follow a "minimal adequate" modeling approach.

An "alternative" to significance-based variable selection is deductive modeling or "effect estimation", as suggested by Harrell (2015: 98), for instance: "[b]y effect estimation is meant point and interval estimation of differences in properties of the responses between two or more settings of some predictors". This method implies pre-selecting variables of interest on the basis of theory and previous research and then reading them off the model as it is generated. Thus, a deductive statistical model includes all theoretically relevant factors and provides information on the effect of each factor, especially the direction and size of the effect. Further reduction of the model is not desirable in this approach, since eliminating a factor from the model would be to eliminate the information on its effect. This means that studies using this approach do not aim for a parsimonious model, or as such for predictive accuracy. Instead, they explicitly assess relevant factors of a variation. They follow Agresti's (2002) advice:

> It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may help reduce bias in estimated effects of other predictors and may make it possible to compare results with other studies. (Agresti 2002: 214)

We believe that deductive modeling is in keeping with the aims of many contemporary linguistic studies: it is a theoretically informed strategy that takes background knowledge and research objectives into account, and it requires the researcher to be clear about their aims and expectations.

Drawing on a random selection of some of our favorite linguistic research papers, we observe that most apply some form of significance-based variable selection and they are more or less explicit about it. Studies such as Lohmann (2011), Fonteyn and Van de Pol (2016), Kaatari (2016) or Hilpert and Saavedra (2020), for instance, report minimal adequate models but do not offer much detail on the procedure followed for variable selection. Some of them do provide model comparison between the minimal and a "saturated" (Kaatari 2016: 548) or full

model, including all variables, by means of the Akaike Information Criterion (AIC). Other papers, such as Wolk et al. (2013), Rosemeyer (2016), Pijpops and Speelman (2017) or Levshina (2016) are more precise about their procedure. Using mixed-effects logistic regression, Wolk et al. (2013) explain their backward selection method:

> [f]irst, we constructed models containing all predictors and all putatively relevant interactions. These models were then reduced by removing predictors and interactions that did not have reliable effects, and the new models were compared to the fuller ones by means of the Akaike Information Criterion. (Wolk et al. 2013: 16)

Rosemeyer (2016: 19) states the use of a backward fitting process based on ANOVA and reports *C* and AIC scores to show that no explanatory power (in the statistical sense) is lost by excluding variables. Pijpops and Speelman (2017: 227) differentiate between "hypothesis-driven" and "nuisance" variables, thus stating the objectives for the variables (where "nuisance" roughly corresponds to "exploration" above). After running a bidirectional stepwise selection, all their hypothesis-driven variables make it into the final model – however, this selection procedure does not pay heed to the initial distinction between variables, and if the outcome had been different (i.e., one or more hypothesis-driven variables dropped), the researchers could not have reported all the relevant results. Last, taking a Bayesian approach, Levshina (2016: 253) reports a model that "contains all 17 variables of interest as fixed effects" – so essentially a deductive model – and notes that a "more parsimonious model with only those predictors whose 95% credible intervals do not include zero […] reveals highly similar results".

Overall, we see that most cowboys – especially those with training and experience in the use of statistical methods – are aware of the fact that their cameras need adjustment, and that the automatic filter is to be used with caution. However, linguists who are new to statistical methods might be misled into thinking that generating minimal models is the one (and only) standard procedure that grants validity and reliability. Some misconceptions might also arise that minimal modeling provides some kind of advantage when dealing with unruly data (e.g., multicollinearity, scarcity of data points, empty cells in the analysis). As pointed out by Harrell (2015),

> [f]or reasons of developing a concise model or because of a fear of collinearity or of a false belief that it is not legitimate to include "insignificant" regression coefficients when presenting results to the intended audience, stepwise variable selection is very commonly employed. Variable selection is used when the analyst is faced with a series of potential predictors but does not have (or use) the necessary subject matter knowledge to enable her to prespecify the "important" variables to include in the model. (Harrell 2015: 67)

Importantly, "important" refers to theoretical, not statistical, importance here. This importance motivates the inclusion of a variable (except perhaps for purely exploratory variables about which no "subject matter knowledge" is present). With minimal modeling, important information, which was originally in our data, might be left out of a model due to the categorical character of variable selection: a predictor either stays in the model or is dropped. Modeled effects, however, are gradual – coefficients, standard errors, *t*-/*p*-values are all continuous measures – so a categorical in/out decision may cause us to lose relevant detail. Moreover, it can be a form of cherry-picking by statistical significance, as Harrell (2015: 63) explains: "Variable selection is an example where the analysis is systematically tilted in one's favor by directly selecting variables on the basis of *p*-values of interest, and all elements of the final result (including regression coefficients and *p*-values) are biased". This bias can lead to overrating the statistical significance of the effects that remain in the model (cf. Heinze et al. 2018: 435).

Minimal models also hinder comparability: when a model is applied to new data, the two results may be different after the variable selection process. Because models are inherently dependent on the observed data, two minimal models including the same candidate variables but run on separate datasets may contain different variables on the basis of their contribution to explaining the variance in the dependent variable. In the camera metaphor, variable selection might make it less straightforward to look back at the two pictures to compare them and find differences as regards perspectives or shades of color. Assessing such differences can be part of a research design or of a research cycle to create cumulative knowledge (cf. Schmidt 1996). In this paper, we argue that a pre-defined set of variables might be useful when these variables are tested on different dependent variables and then compared (e.g., different loci of phonetic reduction in the same morphophonological unit).

1.3 Deductive versus predictive modeling: a matter of research design

There seems to be little awareness in the (linguistic) research community of the distinction between "explanatory" and "predictive" modeling (cf. Shmueli 2010). On the one hand, explanatory modeling applies "statistical models to data for testing causal hypotheses about theoretical constructs". On the other hand, predictive modeling applies "a statistical model or data mining algorithm to data for the purpose of predicting new or future observations" (Shmueli 2010: 291). These definitions of explanatory and predictive modeling refer to the purpose of modeling (not the method). The kind of research setting we address here, empirical

research based on cognitive or social theories of language, can be considered "theory-heavy" (Shmueli 2010: 290). The questions asked in this setting typically call for an explanatory modeling approach, as they aim to test or refine theoretical constructs.[3] This is why it is central to distinguish between these two different dimensions (explanatory vs. predictive) in order to build a model that best "fits" the data as well as the purpose: we should consider that "[e]xplanatory power and predictive accuracy are different qualities; a model will possess some level of each" (Shmueli 2010: 305). One purpose of the present paper is to highlight the distinction between predictive and explanatory modeling and how they lead to a minimal or deductive use of variables, respectively.

As an initial step, both approaches require some kind of pre-selection of candidate variables. In the research setting described above, this will ideally be based on subject matter knowledge (theoretical and empirical literature, domain expertise). Then, a predictive approach is interested only in information that will reliably "predict" an outcome, typically producing a "minimal adequate" model.[4] Thus, predictive modeling entails that some of the variables we or others have hypothesized to have an effect on a given outcome might simply disappear because we are interested in forecasting future results rather than in comprehensively accounting for the existent effects. The explanatory approach, however, requires a careful formulation of expectations as regards the effects of given variables (in relation to their status as in Section 1.2.), and then to estimate them as precisely as possible. This is achieved by a deductive modeling strategy that does not exclude "inefficient" variables. The fact that prediction implies economy of predictors should not be used to argue that deductive models are "uneconomic", undiscerning, or unreliable. It is the researcher's responsibility to craft deductive models that allow for a useful interpretation of effects in a dataset. There are ways to attain this, and there are advantages to this approach.

In what follows, we show the ramifications of both the minimal and deductive approaches to regression modeling by considering their applications to data from a previous corpus study. This case study illustrates some advantages of deductive models: they rely explicitly on prior knowledge, they are responsive to our linguistic objectives, and they allow for estimation and comparison across datasets (in this case, different loci of variation).

---

3 Other areas, e.g., computational linguistics, may be more data-driven and work in an "algorithmic modeling culture" (Breiman 2001) that explicitly prioritizes prediction over explanation. Most of what we suggest here will not apply to the latter approach.

4 Other methods than variable selection have been developed, such as weighting and penalization of coefficients (e.g., 'elastic net'; cf. Tomaschek et al. 2018) – these are designed for predictive modeling, but might be of use in explanatory settings, too (for example to deal with collinearity problems).

2 Case study

To illustrate the outcome and interpretation that different approaches produce, we use data from Lorenz and Tizón-Couto (2017). This study investigated pronunciation variation in English semi-modals, based on the *Santa Barbara Corpus of Spoken American English* (SBC) (Du Bois et al. 2000). We focus here on 337 tokens of *have to*. Their phonetic form was analyzed by the variation in three sounds:

- The final vowel of *to* as /ʊ/ versus /ə/. Reduction to schwa can occur in any instance of *to* and is not expected to be specific to modal items or *have to*.
- Lenition of /t/ to /ɾ/ (or even to zero). /t/-flapping is common in American English, though not at the onset of words or morphemes (cf. Patterson and Connine 2001); therefore, when /t/-lenition occurs in *have to*, it constitutes strong phonetic reduction and attests to the coalescence of the bigram into a single unit.
- Fricative devoicing: /f/ for /v/. In *have to*, the fricative may be devoiced in assimilation to the following /t/. This indicates a degree of coalescence of *have* + *to*, but word-final fricatives are often devoiced (cf. Shockey 2003: 30) and this does not constitute phonetic reduction.

We wanted to see how each of these variations is affected by a number of factors of speech reduction. These are factors for which we have an expectation as to their effect on reduction, following from the literature on variation in speech. In terms of the types of objectives introduced above (Section 1.2), we could label these factors "confirmatory" if the question was only whether or not they show an effect in the expected direction; but since we also want to assess and compare the size of their effects, our aim is "estimation". They are as follows:

- Speech rate: The pace of articulation in the intonation unit (excluding the target item), measured in syllables per second (syll/sec). It is expected that rapid speech increases the chance of unplanned, articulatory reduction (Fosler-Lussier and Morgan 1999; Raymond et al. 2006). The variable is logarithmized and centered in the models below.
- Following sound: The first phoneme of the next word after *have to* (usually the infinitive verb), grouped by place of articulation: labial/dental, alveolar, velar, and "other" (vowel/pause/end of utterance). The "other" category comprises environments in which reduction is known to be less likely (cf. Fox Tree and Clark 1997; Jurafsky et al. 1998; Raymond et al. 2006).

- Stress accent: a light or heavy prosodic stress accent on the main verb (*have*). Items with a heavy stress accent are expected to undergo less reduction (cf. Greenberg et al. 2002; Raymond et al. 2006).
- Speech situation: Based on the SBC file descriptions, situations are private conversations, professional interactions, or public talks. Reduced pronunciation variants are often marked for informality and therefore more expectable in private settings.
- Speaker's year of birth: ranging from 1903 to 1980. If a reduced form takes hold as a variant, it will be increasingly frequent in younger speakers. The variable is scaled and centered in the models below.

As we are estimating the effects of these factors on the phonetic variation in three different positions ([v,f], [t,ɾ], [ʊ,ə]), we employ three models, each with a different dependent variable and the same five factors as independent variables.[5] The three models are built from the same set of tokens of *have to*. Importantly, we need to compare the outcomes in two ways: (a) For each variation, what factors determine it? (b) For each factor, how does it affect each variation?[6]

We now have two possible approaches available for these models, that is, either we try to arrive at "minimal adequate models" by variable selection, or we use "deductive modeling" and consider the effect size of every factor in every model. We presented deductive models in the original study (Lorenz and Tizón-Couto 2017).[7] In the present exercise, we will compare the models and results under either approach; discussing the advantages and disadvantages, we will make a case for the (hitherto) lesser-used option, deductive modeling.

5 There is no problematic collinearity between any of the independent variables (*vif* < 2 for all terms).

6 A reviewer suggested to make this comparison within a single model, by implementing the different variational positions as a moderator variable and then reading the differences from the interaction terms of 'position' with the other variables (see Gahl and Baayen 2019 or Lorenz 2020 for examples). Such an analysis is possible with our data, and it can be deductive. There are also promising methods that can handle multiple dependent variables in one model, notably structural equation modeling (Larsson et al. 2020). We present three separate models here because the three variations are different phonetic phenomena, and because for didactic purposes we want to show the comparison of separate models.

7 For demonstration purposes, we will take a 'strictly deductive' approach here, limiting the analysis to main effects and matching them against the expectations formulated beforehand. In the original study, we had explored interaction effects as well, leading in part to more detailed results.

Table 1: "Full model" of fricative devoicing (Model 1).

	Coeff.	S.E.	Wald Z	*p*	Sig.
Intercept	0.775	0.288	2.69	0.007	**
Speech rate	−1.312	0.451	−2.91	0.004	**
Stress_accent=heavy	1.235	0.25	4.94	<0.001	***
Year_of_birth	−0.114	0.129	−0.89	0.375	
Following_sound=labial/dental	0.284	0.31	0.92	0.359	
Following_sound=velar	0.205	0.38	0.54	0.590	
Following_sound=other	0.475	0.342	1.39	0.165	
Situation=professional	−0.007	0.294	−0.03	0.980	
Situation=public	−0.289	0.51	−0.57	0.572	
	p (chi^2) < 0.001; C = 0.716; D_{xy} = 0.431; AIC = 422.0				

2.1 Deductive models versus backward variable selection

We begin with the model for fricative devoicing ([v] versus [f]) in *have to*. With the deductive approach, we include the five independent variables above. The resulting full model is presented in Table 1.[8]

Trained in "star-gazing" (McElreath 2016: 167), our attention will turn to the rightmost column to scan for significant effects. But richer information is gained from the coefficients and their standard errors. These provide the effect estimation that the model is designed for. Coefficients are odds ratios on the logarithmic scale, which is not the most intuitive of concepts, but their interpretation is straightforward (see, e.g., Jaccard 2001: 7–8 for a concise explanation of log odds ratios). Positive values indicate positive effects (in this case, on the probability of devoicing), negative values indicate negative effects; large values indicate large effects, small values indicate small effects.[9] The standard error (S.E.) shows the precision with which the effect is measured in the data (the smaller S.E., the more precise). *p*-values and significance thresholds, on the other hand, do not inform us of the direction, size, and precision of the effect. An effect may be "significant" because it is large, or because it is precise, or both.

8 Naturally, this "full model" is not the largest model possible (as we could include more variables or interactions); we call it 'full' in opposition to the 'minimal adequate model'. Heinze et al. (2018) call it the 'global model'.

9 We should keep in mind that the coefficients of categorical and continuous variables cannot be directly compared – for categorical variables they refer to factor levels, for continuous variables to an increment of one (i.e., a difference of 1 S.D. = 15 years in YEAR OF BIRTH, or exp(1) = 2.72 syll/sec in SPEECH RATE).

In estimating effects based on coefficients and standard errors, we acknowledge what the data at hand reveal (the coefficient) and apprehend the degree of uncertainty (the standard error) that stems from the fact that the data are only a sample. Standard errors often reflect the amount of data that an effect is measured on. In Table 1, the largest S.E. is for SITUATION=public, because there are only 22 tokens of this factor level; the observed effect is in the expected direction (less devoicing, i.e., more canonical forms, in public speeches), but given the low number of data points, the estimate comes with a high degree of uncertainty.

As we now take the "parsimonious modeling" approach, we will reduce the model in Table 1 to a "minimal adequate model" by statistical variable selection. There are several methods and tools available for this (see Levshina 2015: 149–152 for an overview). We will use stepwise backward variable selection with the R functions `fastbw()` (package `rms` [Harrell 2017]) and `step()` (based on `drop1()`, from the package stats, which comes with the basic installation of R). Stepwise backward variable selection methods start from the full model and take out the least predictive variable; this is repeated until the resulting model would be significantly "worse" (in terms of predictive power) than the previous one. Both `fastbw()` and `step()` apply the Akaike Information Criterion (AIC) for model comparison, though `step()` tends to be more conservative than `fastbw()`. The resulting minimal models are shown in Table 2 (`step()`) and Table 3 (`fastbw()`).

A first essential insight is that there is not *the* one "minimal adequate model". The different selection functions produce different results. Model 2 retains SPEECH

Table 2: 'Minimal adequate model' of fricative devoicing according to `drop1()/step()` (Model 2).

	Coeff.	S.E.	Wald Z	*p*	Sig.
Intercept	−0.574	0.180	−3.19	0.001	**
Speech rate	−1.285	0.440	−2.92	0.004	**
Stress_accent=heavy	1.257	0.241	5.21	<0.001	***
			p (chi^2) < 0.001; C = 0.710; D_{xy} = 0.420; AIC = 413.1		

Table 3: 'Minimal adequate model' of fricative devoicing according to `fastbw()` (Model 3).

	Coeff.	S.E.	Wald *Z*	*p*	Sig.
Intercept	−0.641	0.177	−3.63	<0.001	***
Stress_accent=heavy	1.375	0.236	5.82	<0.001	***
			p (chi^2) < 0.001; C = 0.663; D_{xy} = 0.326; AIC = 420.1		

RATE along with STRESS ACCENT, Model 3 drops it. Having to choose between models, the researchers are forced to make a decision that they may have hoped to leave to automatized statistical procedures. We might resolve this by a principled decision for the "most minimal" (Model 3) or for the smallest AIC (Model 2), or – ideally – by smallest AIC and ANOVA comparison (which in this case suggests that Model 2 is a significantly better fit). Yet, the case serves to show that trying to obtain the most "objective" result by running a variable selection function is deceptive – human decisions are still involved.

Comparing the full model (Model 1) and Model 2, we see that Model 2 has a lower AIC and almost equal *C*, so, if we are aiming for parsimony, Model 2 wins. We would probably even report the same basic findings from either model, such as a higher probability of devoicing in slow speech and with heavy stress accent. This is reassuring, as it shows that results – at least when they are clear enough – will not be turned on their head by changing one methodological decision. However, the deductive full model and a focus on coefficients and standard errors allow us to quantify and assess "significant" and "non-significant" effects alike, rather than merely reporting them as present or absent.

Moreover, there is the issue of comparability: the effect of a factor can be compared across cases when models with the same variables are applied. We will do this by looking at the same factors for the realization of /t/ (full or lenited) and of the final vowel ([ʊ] or [ə]) in *have to*. As above, we present for each variation the full model and "minimal adequate models" as produced by `drop1()/step()` and `fastbw()`.

Overall, the results in Table 4 are in line with what we would expect in a case of articulatory reduction, such that reduction is disfavored in formal situations, and before pauses as well as (slightly) in slow speech – while the effect of STRESS ACCENT cannot be ascertained. If we glance over the models, we find, again, that the estimates of the reported effects do not differ much between the full and minimal models. The difference is in what the minimal models leave out; most strikingly, `fastbw()` drops the variable SITUATION, which produces a significant effect in the other models (/t/-lenition is less likely in professional situations compared to private conversations). So the choice of a variable selection method would lead to either stating that the situational context is a determinant of /t/-lenition or that it has "no effect" – two opposite conclusions, drawn from the same dataset. What we rather want is to quantify the effects and assess their reliability. We see from the full model that the effect of SITUATION=professional matches our expectation (as casual situations favor reduction) with a fair degree of certainty; the effect of SITUATION=public actually shows the expected direction but cannot be reliably ascertained as it is only a very slight trend on a small number of observations. This is the information we ask for in an "effect estimation" scenario.

Table 4: Full model of /t/-lenition, and "minimal adequate" models according to `drop1()/step()` and `fastbw()`.

	Full model				Drop1/step				Fastbw			
	Coeff.	**S.E.**	***p***	**sig.**	**Coeff.**	**S.E.**	***p***	**sig.**	**Coeff.**	**S.E.**	***p***	**sig.**
Intercept	−1.15	0.36	0.001	**	−1.26	0.31	<0.001	***	−1.61	0.29	<0.001	***
Speech rate	1.13	0.62	0.065	.	1.28	0.60	0.033	*	1.44	0.59	0.014	*
Stress_accent=heavy	−0.29	0.36	0.429		Dropped				Dropped			
Year_of_birth.c	0.15	0.19	0.447		Dropped				Dropped			
Following_sound=labial/dental	−0.68	0.44	0.116		−0.70	0.44	0.109		−0.57	0.43	0.182	
Following_sound=velar	0.10	0.46	0.819		0.10	0.45	0.819		0.24	0.44	0.579	
Following_sound=other	−1.71	0.67	0.010	*	−1.78	0.66	0.007	**	−1.61	0.66	0.014	*
Situation=professional	−1.52	0.63	0.015	*	−1.54	0.63	0.014	*	Dropped			
Situation=public	−0.23	0.82	0.782		−0.48	0.79	0.542					
	p (chi^2) < 0.001; C = 0.742; D_{xy} = 0.483; AIC = 241.0				p (chi^2) < 0.001; C = 0.737; D_{xy} = 0.474; AIC = 238.3				p (chi^2) = 0.001; C = 0.695; D_{xy} = 0.391; AIC = 242.8			

In fact, when SITUATION is taken out, the estimates for other effects also change – SPEECH RATE has a larger coefficient, and to the strict "star-gazer" it may look as though FOLLOWING SOUND=other has "lost" one star. These changes are not dramatic when considered with care, but the danger is real: "it may happen, that after eliminating a potential confounder another adjustment variable's coefficients moves closer to zero, changing from 'significant' to 'nonsignificant' and hence leading to the elimination of that variable in a later step" (Heinze and Dunkler 2017: 8).

Regarding the variation in the final vowel (Table 5), the variable selection functions both retain only the factor FOLLOWING SOUND, and the only significant effect is for the labial/dental level (which disfavors reduction to schwa). This is also quite clear from the full model. We might be tempted to say that this time variable selection really only clears the model from uninformative clutter. But let's recall that the aim of variable selection – at least theoretically – is predictive modeling, that is, constructing a model that can optimally predict new data (and does not overestimate spurious effects in the present data). A measure for predictive power is the *C*-index, and with $C = 0.611$, the minimal model in Table 5 is clearly not good at predicting.[10] (Neither is the full model, of course.) This corresponds to the different inherent purposes of the approaches: The full model allows us to estimate the effects of a set of pre-selected variables, and in this case we find that most of them show no reliable association with the dependent variable; with the "minimal adequate model" we try to make predictions about the probability of variants, and in this case we have to admit defeat – the minimal model is not adequate.

2.2 Comparability across models and visualization

We have stated that one motivation for keeping all the (theoretically relevant) variables in our models is their direct comparability across variations. In the present case, we can compare the effect of each factor on fricative devoicing, /t/-lenition and final vowel reduction in *have to*. We illustrate this by visualizing the coefficients and standard errors of the three full models (Figure 2).[11]

10 As a concordance index, *C* technically only tests goodness of fit, i.e., the model's predictions on the same dataset that it was fitted on. Incidentally, the model in Table 5 is also the only one presented here that does not pass the Hosmer-Lemeshow goodness-of-fit test (cf. Hosmer et al. 2013: 157–169).

11 A neat R function for coefficient plots of individual models is `coefplot()` from the package arm (Gelman and Su 2016). To include three models at once, we have extracted the coefficients and S.E. s and plotted them with `ggplot()` (package `ggplot2`; Wickham 2016).

Table 5: Full model of final vowel reduction ([ʊ] or [ə]) in *have to*, and "minimal adequate" model according to `drop1()/step()` and `fastbw()`.

	full model				drop1/step/fastbw			
	Coef	**S.E.**	***p***	**sig.**	**Coef**	**S.E.**	***p***	**sig.**
Intercept	0.51	0.28	0.065	.	0.58	0.22	0.008	**
Speech rate	0.52	0.42	0.213			Dropped		
Stress_accent=heavy	0.09	0.24	0.707			Dropped		
year_of_birth.c	0.04	0.12	0.728			Dropped		
Following_sound=labial/dental	−1.04	0.30	0.001	***	−1.02	0.29	0.001	***
Following_sound=velar	−0.13	0.36	0.715		−0.08	0.36	0.822	
Following_sound=other	−0.08	0.33	0.819		−0.06	0.32	0.845	
Situation=professional	0.11	0.28	0.692			Dropped		
Situation=public	0.24	0.50	0.627					
	p (chi^2) < 0.017; C = 0.630; D_{xy} = 0.261; AIC = 448.9				p (chi^2) < 0.001; C = 0.611; D_{xy} = 0.221; AIC = 440.8			

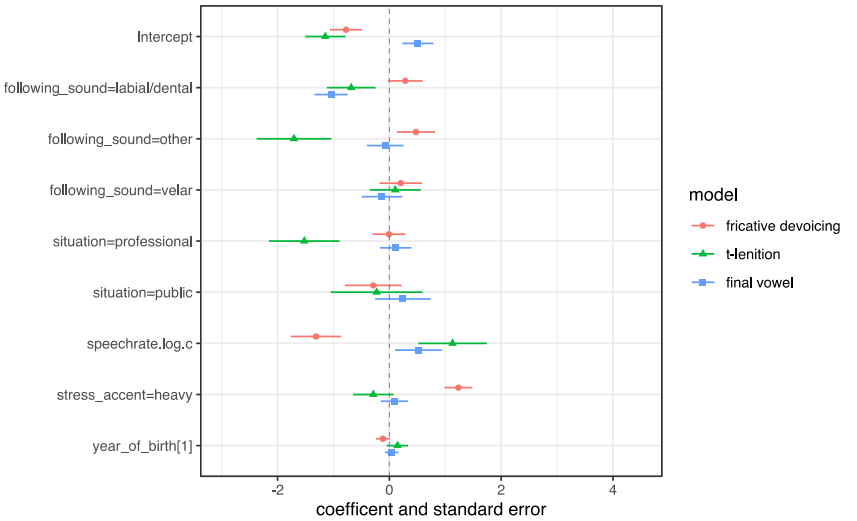

Figure 2: Comparison of effects: Coefficients and standard errors of the full models.

Figure 2 contains all the information needed to interpret the variations at hand. Effects are plotted on the *x*-axis: Negative effects show to the left of the zero line, positive ones to the right; the distance from zero indicates the size of the effect, the error bars its degree of uncertainty. While each color/shape represents the model for one variation, the effects are grouped by factor (level). With this, we can compare how the effects differ between variations in direction and size. For example, we can identify factors that show a clear effect on variation but not the others (e.g., SITUATION=professional), as well as those which barely have any effect at all (e.g., FOLLOWING SOUND=velar); and we can observe that effects on fricative devoicing (red dots) and /t/-lenition (green triangles) tend to pull in opposite directions (e.g., for FOLLOWING SOUND=other, SPEECH RATE, STRESS ACCENT). This is of interest as it shows that there is a variant "haf to/hafta" as opposed to an articulatory reduced form "havda".

Reading off the effects from a coefficient plot like Figure 2 is convenient for a general comparison. For more detail, they can be visualized as in Figure 3, where the estimated effects of speech rate are plotted against the dependent variables.[12]

12 The plots in Figure 3 were created with the R package visreg (Breheny and Burchett 2017). They are partial effect plots, that is, they show the effect of SPEECH RATE assuming all other variables take their mean or most frequent value (which is FOLLOWING SOUND = alveolar, STRESS ACCENT = heavy, YEAR OF BIRTH = 1956, SITUATION = private).

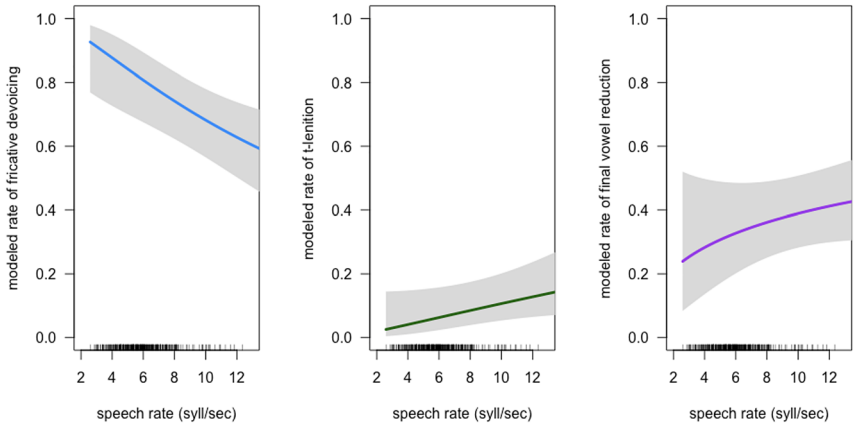

Figure 3: Effects of SPEECH RATE on fricative devoicing, /t/-lenition and final vowel reduction in *have to*.

While higher speech rates increase the odds for /t/-lenition (middle panel), fricative devoicing shows the opposite effect (less devoicing in faster speech, left panel); and there is a slight tendency towards final vowel reduction in rapid speech (right panel), an effect we might have expected but that is weak (and non-significant) in the model. Also, the rate of /t/-lenition overall is lower than that of fricative devoicing or final vowel reduction, which is apparent in the positions of the regression lines. Again, this direct comparison is only possible if the variable in question (here, SPEECH RATE) is included in each model, and it is most reliable when each model includes exactly the same set of independent variables. These conditions are met with pre-specified, "deductive" models but not with "minimal adequate" models.

What if we tried comparisons like those in Figures 2 and 3 with models that have undergone statistical variable selection? To illustrate the difference, we will show this with the more radically trimmed-down models, those suggested by `fastbw()`. The effects plot, showing the coefficients and standard errors from the three models, is presented in Figure 4; the effects can be read in the same way as in Figure 2. As for variable selection, we can only note the presence and absence of effects in Figure 4 (e.g., no effects on fricative devoicing for FOLLOWING SOUND, or on the final vowel for SPEECH RATE and STRESS ACCENT); the variables SITUATION and YEAR OF BIRTH are not even shown, as they have been eliminated from all three models. Thus, instead of showing an accurate comparison, we have turned some effects into a yes/no question – we have given up information for the sake of parsimony. Even if we wanted to show only "significant" effects, this is not achieved, as the coefficient of an individual factor level can still be close to zero (e.g., FOLLOWING SOUND=velar).

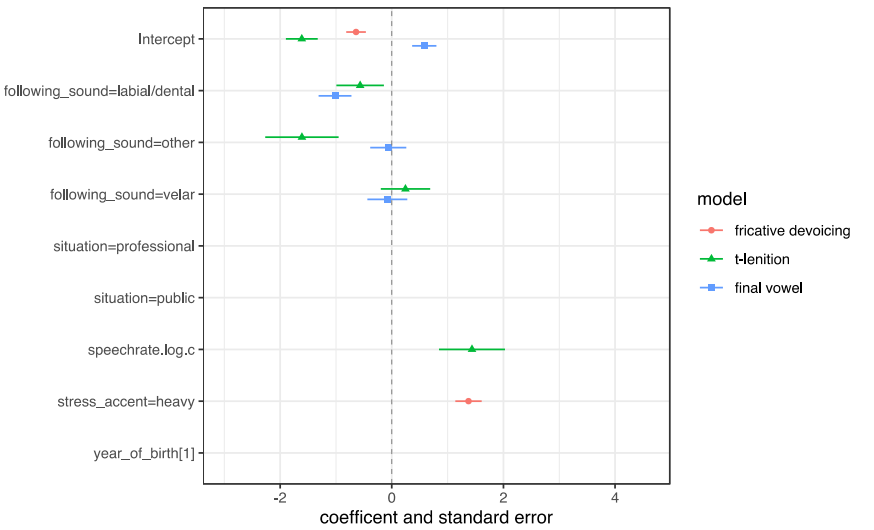

Figure 4: Coefficients and standard errors of the minimal models.

The logic of eliminating variables is based on null hypothesis testing. The observed data do not allow us to reject a null hypothesis, for example that there is no effect of SITUATION on /t/-lenition, with sufficient certainty. We end up excluding the variable, and make no statement about its effect – not even a statement on its effect size being rather too small or its variance too large due to small token numbers. As Cumming (2012: 8) notes, "[s]uch dichotomous decision making seems likely to prompt *dichotomous thinking*, which is a tendency to see the world in an either-or way" (emphasis in original). A variable either contributes to explaining the distribution in the dependent variable, or not. What we should promote instead (according to Cumming 2012, and we agree) is *estimation thinking*, i.e., a focus on effect sizes. The question then is not "yes/no" but "how much". Again, it is the full models that have this very focus.

On the other hand, with the minimal models of Figure 4, we are probably inclined to make some "how much" statements too, e.g., that the effect of SPEECH RATE on /t/-lenition is quite large. Apart from the potential issue of accuracy (cf. Harrell 2015: 69–70), this means that we have taken two steps: First, we test variables on a "yes/no"-question (does it make a significant contribution to the model?); then, if the answer is "yes", we ask "how much" (how large and in what direction is its effect?). It seems that the first step is unnecessary. Compare this to the full models (as seen in Figure 2), where "how much" is all we ask. The answer for some effects may be "practically zero", which is tantamount to "no effect"; but

it still is an answer that can be specified (by coefficient and standard error) and compared to other effect sizes.

3 Deductive and minimal models: summary and discussion

To sum it up, the advantages of a deductive modeling approach are:
- Its groundedness in theory;
- The potentially higher accuracy of the coefficients;
- The comparability of effects across models.

When choosing a deductive modeling approach, we prioritize these points over parsimony and concerns for prediction.

"Groundedness in theory" means that the method takes as its basis the researcher's thoughts and ideas about a question regarding our understanding of language and language use. This is not limited to (dis)confirming existing theories, as deductive modeling can be used to test tenuous theoretical claims and new ideas, too. As for accuracy of coefficients, this can be compromised also by issues arising from the data (e.g., scarcity of certain factor levels, collinearity), especially with "uncontrolled" data from corpora. This means that coefficients should always be read with care and with a mind for the linguistic patterns that produce the observed effects.

Regarding the point of comparability, our example has focused on comparing realizations of different sounds within the same item which can undergo phonetic adaptations that relate to ease of articulation. In other words, we have compared a fixed set of independent variables across different dependent variables, in the same dataset. Another important type of comparison is when the same factors are tested on new data (e.g., from a different but comparable corpus), basically replicating a model. This can help refine effect estimates and reduce uncertainty, leading to cumulative knowledge construction. A breakdown into significant versus discarded factors is of limited usefulness to this goal. Cumming (2012) and Thompson (2002) make this argument in more detail under the keyword of "meta-analytic thinking".[13]

13 Cumming defines "meta-analytic thinking" as "the application of estimation thinking to more than a single study" (2012: 9) in order to reduce the uncertainties of a single data sample. While our case study shows some meta-analytic thinking – in deriving its variables from previous research and in comparing them across variations – it is slightly different from what Cumming (2012) has in mind (see also Thompson 2002).

Under a deductive modeling approach, we are interested in testing the effects of influencing factors on the dependent variable, rather than in predicting outcomes. We think that this motivation underlies most uses of multivariate statistical modeling in linguistic research. Thus, we want to make statements about effects, such as "fricative devoicing in *have to* is less likely at higher speech rates" – and we can specify further how much of a difference speech rate makes: at 4.8 syll/sec (1st quartile) the modeled rate of devoicing is 0.60 (95% CI [0.52, 0.66]), at 6.9 syll/sec (3rd quartile) it is 0.48 ([0.41, 0.55]).[14] This quantifies a specific effect (in Model 1 above). It is more informative to our research questions than prediction statements such as "the probability of fricative devoicing in a token of *have to* at a speech rate of 4.8 syllables per second, with a heavy stress accent is 0.72" (as would be a prediction from Model 2 above). In other words, our purpose is not to give an exact and complete description of the dependent variable (i.e., under what circumstances which pronunciation of *have to* will occur); our purpose is to understand the influence of given variables for which we have hypotheses on the theoretical level (i.e., to assess articulatory reduction vs entrenched pronunciation variants). Statistical models are a tool, and when using a tool it is important to keep the purpose in mind. Here we side with Egbert et al. (2020: 42), who submit that "statistical tests cannot replace linguistic analysis; they are, and should remain, tools that assist the researcher in drawing linguistically valid conclusions".

This is not to say that statistical variable selection is "wrong" – it is a clever but automatic feature (auto-focus) of the "camera" of regression modeling that we can choose to apply, or not. If the purpose is predictive or exploratory, it may be advantageous to use them. However, we can see a number of possible motivations for employing variable selection algorithms that really are not good reasons. We list them here, each with a rebuttal.

I. "The dataset is too small, so the number of model coefficients needs to be reduced". The problem of too few tokens for too many variables cannot be solved by statistical methods of model reduction, as the required number of tokens must be based on a full model with all candidate variables (cf. Heinze and Dunkler 2017: 7–8). Within a regression framework, the only solution is either more data or a tighter pre-selection based on theoretical considerations, research objectives, and/or priorities.

II. "Eliminating some variables reduces the complexity of the reported results". When studying complex phenomena, seeking simplicity can be misleading and does not do justice to the object of study. Rather than simplicity, we should seek clarity. We think that clarity can be achieved by identifying

14 The values are derived from the model with the `Effect()` function from the R package `effects` (Fox 2003).

variables of interest, stating their expected effects, and interpreting the observed results in light of these expectations. Even if they are "null results", discussing them may inform future research and theory-building. Moreover, a "simpler" output comes at the cost of more complex statistical mechanisms: "in search of simpler models, statistical analysis gets actually more complex, as then additional problems such as model instability, the possibility of several equally likely competing models, the problem of postselection inference, etc. has to be tackled" (Heinze et al. 2018: 432).

III. A "minimal adequate model" is more objective because it results from a strict statistical procedure". Research results are never fully objective because they (partly) follow from decisions on how to treat and analyze the data, what variables to consider, etc. Moreover, as we have seen, statistical variable selection is not fully determined either, as different algorithms produce different outcomes. In short, statistical procedures are not to replace the researcher's prudent decision-making.

IV. "The 'full model' should be trimmed down to avoid the risk of overfitting". Overfitting means that the model is so tightly tuned in on the present data that it may make wrong predictions on new data (cf. Speelman 2014: 529). If prediction is explicitly not the goal of modeling, the issue is of less concern. That said, we do want the conclusions drawn from a "deductive" model to be as generally valid as possible, so measures like goodness of fit and model optimism provide valuable information (cf. Harrell 2015: 113–116; Steyerberg 2009: 84) – in the model-as-camera allegory we could say that they show us to what extent the model highlights (and perhaps exaggerates) the contrasts in the picture.[15] A bigger concern to deductive modeling, however, is (multi-) collinearity between independent variables, as it can confound coefficient sizes and hence effect estimations (cf. Shmueli 2010: 299). Therefore, checking on "variance inflation factors" (VIF); (cf. Harrell 2015: 78–79; Levshina 2015: 272) is more essential to deductive modeling than measures of overfitting.

15 "Optimism is defined as true performance minus apparent performance, where true performance refers to the underlying population, and apparent performance refers to the estimated performance in the sample" (Steyerberg 2009: 84). The validate() function in rms (Harrell 2017) quantifies optimism through resampling. In our case study, the full models all show somewhat greater optimism than their minimal counterparts. In the models on fricative devoicing (Tables 1–3), this leads to the corrected D_{xy} of the full model being lower than for the minimal models, though they are all close to D_{xy}(corrected) ≈ 0.4. In the models for the final vowel (Table 5), having a low accuracy to begin with, the corrected D_{xy} drops below 0.2. Again, this means that the models should be interpreted with caution for effect estimation, but are dangerously insufficient for prediction proper.

4 Further considerations

We should admit that the examples we have presented here are relatively simple kinds of models. For one thing, the independent variables were all clearly derived from theoretical hypotheses and motivated by the need to estimate their effect (i.e., of type (iii) by the definitions in Section 1.2). This is what Baayen (2008: 236) calls the "ideal" case. Often, a researcher will want to consider additional factors, whose influence is to be explored without particular expectations (type (i)). These could be subjected to a selection mechanism,[16] although there is no need for this when there is no pressure to achieve a parsimonious model. A possible motive may be that the researcher considers them "nuisance variables" (Pijpops and Speelman 2017) and wants to avoid reporting an unnecessarily large model.

We also did not include any interaction terms in our models. Indeed, testing for interactions may seem at odds with a deductive approach when we have no *a priori* hypothesis about them. Yet, it can follow from the same rationale as keeping a "full model": We have chosen a set of variables based on theory and previous research, now we want to see how they affect the dependent variable in as much precision and detail as possible – including "non-significant" main effects as well as relevant interactions that emerge from the data.

Another limitation of the above demonstration is that we did not present a mixed-effects approach.[17] In principle, though, the same considerations apply to the fixed effects in a mixed model: is the aim to estimate the effects of pre-selected variables or is it to achieve a parsimonious, predictive model? Variable selection is particularly tricky because a random effect may "step in" to capture variance otherwise accounted for by a fixed effect (Barth and Kapatsinski 2018).[18]

Finally, one might wonder if or how tree-based models (recursive partitioning, conditional inference trees, random forests) fit into this discussion, given that they are often described as a handy alternative to regression models. Tree models are good at showing what factor combinations likely produce a given outcome. For this, they may indeed be more intuitive, as is sometimes claimed (e.g., Baayen et al. 2013; but see Gries 2020 for some words of caution). However, they do not provide effect estimates for individual variables. We showed above how such effect

16 Both `step()` and `fastbw()` have parameters to set the scope of variables over which the selection is run (see the respective documentations in R).

17 For corpus linguistics, a compelling urge to consider mixed-effects modeling has been put forward by Gries (2015).

18 In a mixed-effects setting, there is also the question of a 'maximal' or restricted set of random effects. A discussion of this is beyond the scope of this paper – see Barr et al. (2013) and Bates et al. (2015) for different positions.

estimates can be read from a deductive regression model. The nested effects of tree models, however, have to be read like the prediction statement above, i.e., as a probability of a variant choice in a particular constellation of values of the predictor variables. Moreover, tree models inherently involve a strict variable selection (in fact selecting just one variable at every split), so they suffer from the same problems as "minimal adequate" regression models when an explanatory modeling approach is taken.[19] That said, tree models do have their advantages, e.g., robustness to collinearity and data scarcity (cf. Levshina 2021). Again, it is up to the researcher to choose the appropriate tool for the task at hand.

5 Concluding remarks

This paper is intended to raise awareness about how regression modeling can be used as a "camera" to capture linguistic realities. When we analyze linguistic data by means of multivariate statistics, we must be aware that different strategies can lead to different results – so we should be clear about our strategy and not simply assume predictive modeling and "minimal adequate models" as a default. What we have advocated in this article is a deductive modeling strategy that aims at accurate effect estimation rather than parsimony: this approach essentially consists in building a "full" model with carefully selected variables on the basis of linguistic knowledge, previous findings, and research ideas, with no further reduction.

When fitting multivariate statistical models to complex data, we should be mindful of the questions we seek to address, rather than blindly apply an automated procedure that seeks a level of simplicity our linguistic concerns or analysis might not call for. "Pre-selection" of variables by theoretical criteria is advantageous when estimation or even direct comparison of effects is at stake. In this paper, we have provided an example of how putting accuracy before parsimony, by means of deductive modeling, can translate into transparency/clarity as regards (a) research aims and hypotheses, (b) comparability with previous studies, and (c) comparability of models with the same independent variables applied to different dependent variables.

We have also shown that simplicity (parsimony) does not automatically translate into a higher predictive accuracy, especially when the model makes weak predictions to begin with. Dropping variables does not necessarily/always improve model predictiveness, but it does reinforce the researcher's overreliance on

19 When fitting tree models to the data of our case study, the variables they select correspond to the minimal models presented above (depending on parameter settings).

statistical significance levels and brings them closer to a selective reporting of results that might "accidentally" leave out important information (cf. Harrell 2015: 63; Sönning and Werner this issue). The bottom line is that deductive models might also be wrong, but they are certainly helpful in order to estimate the effects in our (more or less) complex data. The theory-driven character of deductive modeling also partly avoids the implication that modeled results provide sweeping predictions that will be corroborated across corpora; such predictions would run counter to the fact that corpora are neither random nor representative samples (cf. Egbert et al. 2020: 14; Koplenig 2019; Leech 2007).

Overall, we think that as a general strategy of statistical inference, the deductive modeling approach can be one element in overcoming the "replication crisis", in particular the "significance bias" of published findings. There is no principled problem with publishing a model that does not predict well (like ours for the final vowel) or even a model that does not contain any significant effects – as long as we clearly and carefully explain our interpretation of these (null) results. In any case, this requires that we have the relevant knowledge to form a judgement about which variables are worth considering on theoretical grounds; linguists who use statistics should first and foremost be good linguists. Or, if we think of models as cameras: The photographer should certainly know how to handle the camera and its settings, but they should never forget to point the camera in the right direction.

Acknowledgments: We would like to thank the participants and audience at the workshop "The 'quantitative crisis', cumulative science and English linguistics" at the ISLE 5 conference for inspiring debates that helped shape the positions presented here. We are especially indebted to Lukas Sönning and Valentin Werner, as the conveners of the workshop and editors of this issue, for more than the usual amount of support and feedback; and to the editor-in-chief of Linguistics, Volker Gast, as well as two anonymous reviewers whose constructive criticism has done much to improve this paper. All remaining errors are entirely our own.
Research funding: The research reported in this article was funded by the Spanish Ministry of Science and Innovation (grant PID2020-118143GA-I00) and Xunta de Galicia (grant ED431C2021/52); support is gratefully acknowledged.

References

Agresti, Alan. 2002. *Categorical data analysis*. Hoboken, NJ: Wiley.
Baayen, R. Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
Baayen, R. Harald. 2013. Multivariate statistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 337–372. Cambridge: Cambridge University Press.

Baayen, Harald R., Laura A. Janda, Tore Nesset, Endresen Anna & Anastasia Makarova. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37(3). 253–291.

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

Barth, Danielle & Vsevolod Kapatsinski. 2018. Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. In Dirk Speelman, Kris Heylens & Dirk Geeraerts (eds.), *Quantitative methods in the humanities and social sciences*, 99–116. Cham: Springer.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2015. Parsimonious mixed models. *ArXiv preprint.* https://arxiv.org/abs/1506.04967v1.

Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Box, George E. P. 1979. Robustness in the strategy of scientific model building. In Robert L. Launer & Graham N. Wilkinson (eds.), *Robustness in statistics*, 201–236. New York: Academic Press.

Breheny, Patrick & Woodrow Burchett. 2017. Visualization of regression models using visreg. *The R Journal* 9(2). 56–71.

Breiman, Leo. 2001. Statistical modeling: The two cultures. *Statistical Science* 16(3). 199–231.

Cumming, Geoff. 2012. *Understanding the new statistics: Effect sizes, confidence intervals and meta-analysis*. New York: Routledge.

Cumming, Geoff & Sue Finch. 2005. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist* 60(2). 170–180.

Du Bois, John W., Wallace Chafe L., Charles Meyer, Sandra Thompson A., Robert Englebretson & Nii Martey. 2000–2005. *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia: Linguistic Data Consortium. www.linguistics.ucsb.edu/research/santa-barbara-corpus (accessed 1 December 2013).

Egbert, Jesse, Tove Larsson & Biber Douglas. 2020. *Doing linguistics with a corpus*. Cambridge: Cambridge University Press.

Figueiredo Filho, Dalson Britto, Ranulfo Paranhos, Enivaldo C. da Rocha, Mariana Batista, José Alexandre da Silva, Jr., Manoel L. Wanderley D. Santos & Jacira Guiro Marino. 2013. When is statistical significance not significant? *Brazilian Political Science Review* 7(1). 31–55.

Fonteyn, Lauren & Nikki van de Pol. 2016. Divide and conquer: The formation and functional dynamics of the modern English *ing*-clause network. *English Language and Linguistics* 20(2). 185–219.

Fosler-Lussier, Eric & Nelson Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication* 29. 137–158.

Fox, John. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software* 8(15). 1–27.

Fox Tree, Jean E. & Herbert H. Clark. 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* 62. 151–167.

Gahl, Susanne & Harald Baayen. 2019. Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. *Journal of Phonetics* 74. 42–54.

Gelman, Andrew & Yu-Sung Su. 2016. *arm: Data analysis using regression and multilevel/hierarchical models*. R package version 1.9-3. Available at: https://CRAN.R-project.org/package=arm.

Glynn, Dylan. 2014. Correspondence Analysis: Exploring data and identifying patterns. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 443–486. Amsterdam & Philadelphia: John Benjamins.

Greenacre, Michael. 2007. *Correspondence analysis in practice*. London: Chapman & Hall.

Greenberg, Steven, Hannah Carvey & Leah Hitchcock. 2002. The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *Proceedings of the International Speech Communication Association Workshop on Prosody and Speech Processing*, 351–354.

Gries, Stefan T. 2013. *Statistics for linguistics with R*. Berlin & Boston: De Gruyter Mouton.

Gries, Stefan T. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.

Gries, Stefan T. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3). 617–647.

Harrell, Frank E. 2015. *Regression modeling strategies*. Cham: Springer.

Harrell, Frank E. 2017. *rms: Regression modeling strategies*. R package version 5.1-1.

Heinze, Georg & Daniela Dunkler. 2017. Five myths about variable selection. *Transplant International* 30. 6–10.

Heinze, Georg, Christine Wallisch & Daniela Dunkler. 2018. Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal* 60. 431–449.

Hilpert, Martin & David Correia Saavedra. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16(2). 393–424.

Hosmer, David W., Lemeshow Stanley & Rodney X. Sturdivant. 2013. *Applied logistic regression*. Chichester: Wiley.

Hothorn, Torsten, Hornik Kurt & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational & Graphical Statistics* 15. 651–674.

Jaccard, James. 2001. *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage.

Janda, Laura A. 2013. Quantitative methods in cognitive linguistics: An introduction. In Laura A. Janda (ed.), *Cognitive linguistics: The quantitative turn*, 1–32. Berlin & Boston: De Gruyter Mouton.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell.

Jurafsky, Daniel, Alan Bell, Eric Fosler-Lussier, Cynthia Girand & William Raymond. 1998. Reduction of English function words in Switchboard. *Proceedings of ICSLP-98* 7. 3111–3114.

Kaatari, Henrik. 2016. Variation across two dimensions: Testing the complexity principle and the uniform information density principle on adjectival data. *English Language and Linguistics* 20(3). 533–558.

Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.

Larsson, Tove, Luke Plonsky & Gregory R. Hancock. 2020. On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*. Advance online publication https://doi.org/10.1515/cllt-2020-0051.

Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 133–149. Amsterdam: Rodopi.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam & Philadelphia: John Benjamins.
Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268.
Levshina, Natalia. 2021. Conditional inference trees and random forests. In Magali Paquot & Stefan T. Gries (eds.), *A practical handbook of corpus linguistics*, 607–640. Cham: Springer.
Lohmann, Arne. 2011. *Help* vs. *help to*: A multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics* 15(3). 499–521.
Lorenz, David. 2020. Converging variations and the emergence of horizontal links: *to*-contraction in American English. In Lotte Sommerer & Elena Smirnova (eds.), *Nodes and networks in diachronic construction grammar*, 243–274. Amsterdam & Philadelphia: John Benjamins.
Lorenz, David & David Tizón-Couto. 2017. Coalescence and contraction of V-*to*-V_{inf} sequences in American English – Evidence from spoken language. *Corpus Linguistics and Linguistic Theory*. Advance online publication. https://doi.org/10.1515/cllt-2015-0067.
McElreath, Richard. 2016. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: CRC Press.
Patterson, David & Cynthia M. Connine. 2001. Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. *Phonetica* 58. 254–275.
Pijpops, Dirk & Dirk Speelman. 2017. Alternating argument constructions of Dutch psychological verbs: A theory-driven corpus investigation. *Folia Linguistica* 51(1). 207–251.
Raymond, William D., Robin Dautricourt & Elizabeth Hume. 2006. Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18. 55–97.
Rosemeyer, Malte. 2016. The development of iterative verbal periphrases in Romance. *Linguistics* 54(2). 235–272.
Sampson, Geoffrey R. 2005. Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics* 10. 10–36.
Sampson, Geoffrey R. 2013. The empirical trend: Ten years on. *International Journal of Corpus Linguistics* 18(2). 281–289.
Schmidt, Frank L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1(2). 115–129.
Shmueli, Galit. 2010. To explain or to predict? *Statistical Science* 25(3). 289–310.
Shockey, Linda. 2003. *Sound patterns of spoken English*. Oxford: Blackwell.
Speelman, Dirk. 2014. Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In Dylan Glynn & Justina A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 487–533. Amsterdam & Philadelphia: John Benjamins.
Steyerberg, Ewout W. 2009. *Clinical prediction models: A practical approach to development, validation, and updating*. Cham: Springer.
Thompson, Bruce. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31(3). 25–32.
Tomaschek, Fabian, Hendrix Peter & R. Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249–267.
Tong, Christopher. 2019. Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician* 73(1). 246–261.
Upton, Graham J. G. 2017. *Categorical data analysis by example*. Hoboken, NJ: Wiley.

Vasishth, Shravan & Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas: Part I. *Language and Linguistics Compass* 10(8). 349–369.
Vittinghof, Eric & Charles E. McCulloch. 2006. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology* 165. 710–718.
Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer.
Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution* 1(1). 7–18.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.
Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC Press.
Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. New York: Springer.