Too many tokens: Modeling lexical effects in corpora

Susanne Gahl
University of California at Berkeley
gahl@berkeley.edu

R. Harald Baayen
University of Tübingen
harald.baayen@uni-tuebingen.de

Abstract Many statistical models of properties of words in texts and in running speech use as their outcome variables properties of word tokens, rather than properties aggregated over word types. Here, we describe methodological and conceptual problems with token-level (or 'token-based') models, including assumption violations of the models typically used, model redundancy, and distortions of parameters estimating lexical effects – i.e. the very target of what such models seek to capture. We show that these problems persist – and in some ways, grow more extreme – when we follow widely-used practices for specifying random effects structure. We do so by comparing models of spoken word duration of homophones in the Switchboard corpus, a dataset that has been discussed and modeled extensively. We discuss possible remedies and recommendations.

Keywords: corpora; token-level analysis; spoken word duration; mixed-effects regression models; generalized additive models; concurvity

1. Introduction

Statistical models of variation in the realization of words have generally taken one of two approaches. The first approach is to aggregate information over word types: In these models, the dependent variables are averages, such as mean latencies, word duration, vowel formant values, letter stroke duration, typing speed, or other properties of words as articulatory, visual, and acoustic events, averaged by word. This is the approach taken in numerous analyses of controlled, balanced data from experiments (e.g. Wright, 2004), as well as in some corpus-based studies, such as Gahl (2008) and Gahl and Baayen (2024). The second approach models properties of individual tokens, such as trial-level response latencies, token duration, and so on. This is the approach taken in most corpus-based analyses (e.g. Bell et al., 2003, 2009; Fosler-Lussier & Morgan, 1999; Gahl et al., 2012; Gries, 2015; J. Hay, 2007; J. B. Hay et al., 2015; Kilbourn-Ceron et al., 2020; Lohmann, 2018b; Seyfarth, 2014; Tanner et al., 2020).

The obvious appeal of token-level analyses lies in the richness of the information that can be considered, such as local speaking rate, the words preceding or following the target, indexical information about talkers and listeners, and so on. Token-level models conform to the sound methodological principle of 'not throwing away data', i.e. avoiding information loss due to aggregation. Another alluring feature of such analyses lies in the sheer size of data sets, which can support nuanced statistical models. On top of these advantages, the influence of token-level analyses is to some degree self-perpetuating, in that new investigations tend to include token-level models so as to enable comparisons to previous work. Token-level models have effectively acquired the status of a gold standard, with type-based models looking like a less sophisticated alternative. The aim of the current study is to point out a set of methodological problems arising with token-level models of unbalanced data sets, such as naturalistic corpus data. We argue that such models introduce biases that, ironically, can get in the way of understanding lexical effects. We are emphatically not claiming that token-level models are to be avoided under all circumstances. Rather, we wish to point out some consequences of token-level modeling that restrict the usefulness of such models to certain types of inquiry. Simplifying our point somewhat: token-level models turn out to be surprisingly ill-suited for modeling what are often simply referred to as 'lexical effects' or 'word-specific' effects. The optimal statistical analysis ultimately depends on one's theory of the domain being modeled, as well as on the goals of the analysis.

In what follows, we argue that the problems with of token-level models (or models sometimes termed 'token-based', in which the observations being modeled constitute tokens, but whose ultimate analytical target are types) include the following:

- **Ballot-box stuffing:** Because token-based regression models will be penalized for the residual of every token, such models will be driven by high-frequency target words. Because deviations of model predictions are punished for every single observation, token-based models can only achieve excellent fit if they work well for high-frequency words.
- **Model redundancy:** 'Ballot-box stuffing' happens not just once, but for every lexical variable included in a token-based model: Tokens of any given high-frequency target word contribute identical sets of values for all type-level properties of the target. These replicates of co-occurring properties ('clones') of each type can render variables redundant, i.e. predictable from other variables in the model, or from the model as a whole, making it difficult or impossible to evaluate the contributions of any individual predictors.
- **Distorted parameter estimates:** 'Ballot-box stuffing', while harmful for the reasons just outlined, might in principle apply evenly across the ranges of lexical variables. However, at the type level, word frequency is correlated with other variables (Baayen, 2011; Frauenfelder et al., 1993; Köhler, 1986; Landauer & Streeter, 1973), and high-frequency words tend to have certain phonological and semantic properties in common. The resulting similarities across high-frequency types (i.e. similarities across 'voters', rather sets of

identical ballots cast by each voter) has the potential to inflate the predictiveness of some variables and underestimate that of others.

- **Assumption violations:** The uneven distribution of observations, and the correlation of frequency with other variables, lead to violations of modeling assumptions, such as normality, independence, and homogeneity of residuals.

Many researchers trust that the problems mentioned so far can be avoided through judicious model specification, specifically by-word random effects in mixed-effects regression models. Random effects model the distribution of token-level observations within type-based clusters (Baayen et al., 2008; Bates, 2005; Pinheiro & Bates, 2000; Quené & van den Bergh, 2008). This prevents clusters of tokens from gaining undue influence over the model as a whole — or so the assumption goes. Here, we demonstrate that the problems just outlined persist when the random effects structure of the models include by-word random effects.

The consequences of analyzing token-level vs. type-level information have not been obvious, in part because any given study typically only reports only token-level or only type-level results, but not both. In the current study, we report both type-level and token-level models of spoken word duration, in order to draw attention to problems with token-based models. We analyze a dataset that has been previously been analyzed and discussed extensively, *viz.* homophone duration in the Switchboard corpus, i.e. the duration of words like *time* and *thyme*. By focusing on an existing data set, we wish to highlight consequences of methodological choices, rather than properties of specific variables or item lists. We illustrate the problems with token-based analyses and discuss possible remedies, concluding with a set of recommendations.

2. Methods

2.1 Data

We analyzed the same publicly available data set described in Gahl (2008, 2009), Gahl and Baayen (2024), and Lohmann (2018a). With the exception of Gahl (2009), these previous analyses of the data set all used type-level models, i.e. modeling word duration averaged over word types. We refrained from collecting or proposing any novel variables or make other changes to the dataset, for the sake of comparability to earlier studies and so as to maintain the focus on methodological issues.

According to Gahl (2008), the initial list of target words for the data set contained all English word forms with at least one non-homographic homophone, based on the CELEX database (Baayen et al., 1995). Word forms with identical spelling and pronunciation (e.g. nouns and verbs spelled *time*) were treated as tokens of the same type. Gahl (2008)'s exclusion criteria excluded the following items: (1) spellings associated with more than one pronunciation, e.g. *tear* (homophonous with *tier* and *tare*); (2) pairs involving function words, such as *in/inn* and *or/ore* and interjections, such as *whoa/woe*; (3) pairs such as *source/sauce* that are homophones in the (Received Pronunciation of British English) CELEX transcriptions, but unlikely to be homophones in the varieties of American English represented in the Switchboard corpus; (4) items containing transcription errors in CELEX; and (5) names of letters in the alphabet. The resulting list contains 409 target types, represented by 79,219 tokens.

The spoken duration of all tokens of these items was extracted from the time-aligned orthographic transcript (Deshmukh et al., 1998) of the Switchboard corpus (Godfrey et al., 1992), a corpus of 240 hours of telephone conversations between strangers. We initially retained tokens that were immediately followed by an unfilled pause (defined as a period of silence of 0.5 seconds or longer) or by a filled pause (e.g. a hesitation marker such as *um, uh, err*) and allowed Fluency, coded as a binary factor, to interact with talker age and with the relative frequency of target and homophone. However, the 'disfluent' tokens were too unevenly distributed to allow meaningful models of these

interactions. Therefore, we excluded them from all further analyses. Excluding the disfluent tokens left 56,024 tokens of 403 distinct types for analysis.

2.2 Variables considered

2.2.1 Type-level predictors

The type-level variables considered here are those that were included in the multiple linear regression model in Gahl (2008), the first analysis of the dataset, and/or subsequent analyses using GAMMs (Gahl & Baayen, 2024; Lohmann, 2018a). The variables for these analyses were based on prior research on spoken word duration, including Bell et al. (2003, 2009), Gahl et al. (2012), Lieberman (1963), Pluymaekers et al. (2006), Shields and Balota (1991), Sorensen et al. (1978), Walsh and Parker (1983), and Warner et al. (2004).

Residualized baseline duration The residuals of a GAM predicting the target's baseline duration (as the sum of the average phone durations) from properties of word forms, following Gahl and Baayen (2024); model summary and visualization of smooths for that model are in Appendix A. The rationale for residualizing baseline duration, rather than using the baseline duration itself, is that the distribution of phonological segments in the English lexicon is to some extent predictable from lexical variables such as Phonological Neighborhood Density and lexical frequency (as shown for Dutch, English, German and French in Dautriche et al., 2017); therefore, estimates of word duration based on by-segment averages do not accomplish what the baseline duration measure is intended to accomplish, which is to control for the inherent duration of the segments (e.g. the duration of a voiceless alveolar stop [t], followed by [aɪ], followed by [m]) as distinct from from word-level properties such as Phonological Neighborhood Density, lexical frequency, and so on.

Biphone probability The average of the target's position-specific biphone probabilities, based on Vitevitch and Luce (2004).

Lemma frequency The log-transformed frequency of the target (e.g. *time* vs. *thyme*), based on the CELEX database (Baayen et al., 1995).

Relative Frequency The (log-transformed) CELEX frequency of the target, divided by the frequency of its homophone (e.g. the frequency of *time*, divided by the frequency of *thyme*). The same variable was used in Lohmann (2018a). The relative frequency is greater than zero for the higher-frequency member of each pair of homophones, and smaller than zero for the lower-frequency member of the pair. The relative frequency entered the models in the form of a tensor product to model the interaction between `Lemma Frequency` and `Relative Frequency`. The rationale for including that interaction is that the higher the target lemma frequency, the more it can exceed its homophone twin in frequency.

Word form A factor identifying the phonological form that homophones have in common, e.g. /taɪm/ for *time* and *thyme* (termed *lexemes* in Gahl, 2008, following Levelt et al., 1999).

Morphological Complexity A binary factor distinguishing morphological simple vs. complex target words, typically third person singular, *-s* plural, or past tense forms, such as *lacks* (vs. *lax*) or *allowed* (vs. *aloud*).

Noun quotient The estimated proportion of nouns among the tokens of a given form (noun vs. verb uses of *time*), based on CELEX (Baayen et al., 1995). Nouns occur in phrase-final position more often than words of other syntactic categories in English and therefore are more likely to undergo phrase-final lengthening. For example, tokens of *thyme* are more likely to be phrase-final than tokens of *time* (which may represent verbs and therefore less likely to be phrase-final). As this variable was bimodally distributed, it was dichotomized, indicating whether the proportion of noun uses of a given form was above vs. below 0.5.

Orthographic regularity According to Gahl (2008), a measure indexing the average probability of a word's graphemes, normalized by the probability of the most probable pronunciation of each grapheme, based on American English grapheme-to-phoneme probabilities (Berndt et al., 1987).

Pause quotient The proportion of tokens of a given target that were immediately followed by a pause.

Phonological Neighborhood Density (PND) The number of words differing from the target word by addition, deletion, or substitution, based on the English Lexicon Project (Balota et al., 2007).

2.2.2 Token-specific predictors

Our token-based models contain several additional variables pertaining to specific tokens of words, as follows:

Age The talker's age.

Bigram probability The word-based bigram probability of the target, given the word following it.

Sex The talker's sex (male or female, as coded in Godfrey et al., 1992).

2.3 Statistical tools and strategies

2.3.1 GAM(M)s

We made use of the Gaussian Location Scale Additive Model, fitting Generalized Additive Models (GAMs) and Generalized Additive Mixed Models (GAMMs), using the packages `mgcv` (Wood, 2011, 2017; Wood, 2004) and `itsadug` (van Rij et al., 2020) in R (R Core Team, 2022). Tutorials and applications of GAMs to linguistic data can be found, for example, in Baayen and Linke (2021), Baayen et al. (2010), Chuang et al. (2021), Sóskuthy (2021), Wieling (2018), and Wieling et al. (2011, 2014). Here, we summarize properties of GA(M)Ms that are critically important to the current study.

The Generalized Additive (Mixed) Model (henceforth GAM except when specifically referring to models with random effects) is a regression model that relaxes two of the assumptions that underlie Linear (Mixed Effects) Regression models, which it otherwise resembles. The first is the linearity assumption, i.e. the assumption that predicted values change at a constant rate across values of the predictor variables. In GAMs, the relationship between predictors and outcome is not assumed to be linear, but is modeled as the combination of two sets of functions. The second is the assumption of equal variance. Gaussian Location-Scale GAMs, which are the types of models we use here, allow the variance, which is assumed to be Gaussian, to change (non-linearly, if necessary) with the predicted mean. That is to say, both mean and variance are modeled as (potentially non-linear) functions of the predictors. The first ('parametric') set of functions resemble linear predictors in linear mixed models (LMMs); this set includes any continuous predictors specified as linear, as well as fixed-effect factors. The second ('non-parametric') set of terms estimates the relationship between predictors and outcome as a potentially nonlinear function, specifically the sum of successively more complex (more "wiggly") basis functions. The coefficients of these functions are determined by a procedure balancing model fit and parsimony. For example, if a predictor is truly linear, a GAM will penalize any nonlinear components to the point of setting their coefficients to zero.

The estimated number and coefficients of the basis functions are not specified ahead of time, but they can be constrained by the researcher. We did so here, in order to counteract artifactual effects of high-frequency words, we limited the number of basis functions in our initial token-based models, to 3 for univariate smooths (i.e. k = 4) and to 4 (i.e. k = 5) for the tensor product of lemma

frequency and frequency ratio. We consider several alternative numbers of basis functions in our discussion of the models in section 3.8.1.

GAMs can model interactions among continuous variables, yielding model estimates of surfaces, which can be visualized by means of contour plots. These plots may be read in the manner of a topographical map, with elevation indicating predicted target duration. In our plots, 'warm' colors (orange to yellow) indicate 'elevated', i.e. longer, predicted duration. 'Cool' colors (green to blue) indicate shorter predicted duration.

GAMs can include Gaussian random effects, analogous to random intercepts and slopes in LMM. Mixed models, i.e. models including fixed and random effects, are very common in many areas of linguistics and psycholinguistics, and we believe most readers are likely to be very familiar with such models. We comment on random effects explicitly here because of their importance in the methodological issue we wish to point out. Typical uses of random effects in research on lexical processing include by-talker and by-item effects, for clustering observations by talker or by item (e.g. words or experimental stimuli), to allow inferences about entire populations of talkers or items. Rather than estimating a predicted mean for each level of a factor, the idea is to estimate the variance about an overall mean. Most studies of lexical factors in pronunciation variation mentioned above employ by-word random intercepts, estimating the variance of a distribution around the model intercept. Many models additionally include random slopes, estimating the variance of the slope of one or more fixed effects, i.e. the variability in the degree to which, for example, different words are lengthened depending on their position in an utterance or depending on the age of the talker, or the degree to which different talkers' pronunciations vary depending on some experimental manipulation or lexical factor.

As mentioned above, GAMs relax the assumption of equal variance that underlies linear regression models. Instead of relying on an assumption of the variance being equal along the range of the predicted outcome, GAMs can include terms for predicting the variance (along with predicting the mean). We capitalize on this property of GAMs here, by modeling the variance in duration as a function of target frequency (a variable at the center of previous analyses of the data) in several of our models.

In light of the fact that our dataset has been studied previously, particularly in type-based models, and given the exploratory nature of our models, we set $\alpha = .0001$.

2.3.2 Model criticism and evaluation

GAMs, like Linear Mixed Models (LMMs), rely on the assumption that the model residuals do not co-vary with the predictors: Model fit should be about equally good along the full ranges of the predictors.

A second assumption underlying GAMs (and LMMs) is that the by-group adjustments (the 'random effects') are Gaussian noise that is supposed to be independent and identically distributed; among other things, this means that the random effects are assumed to be uncorrelated with the other predictors

We used several tools for comparing models to one another and for model criticism.

AIC The Akaike Information Criterion is a measure of model goodness-of-fit penalized for model complexity, i.e. the number of predictors in the model. Put differently: This measure specifies which of two models fitted to the same datapoints is more likely to have generated the data. Decreasing values of AIC indicate better model fit (measured as 2*Log Likelihood), taking into account the number of predictors.

Concurvity of each predictor in the model, i.e. a measure of the degree to which each component of the model can be approximated by a weighted sum of the other model components. If the smooth term can be predicted to a large extent by the other terms, then there is a concurvity problem: High concurvity renders estimates unstable, i.e. liable to change drastically in

response to seemingly minor changes to the model or data. We assessed the concurvity of each term with all remaining the predictors in each model, i.e. x1 = f(x2) + f(x3) + ... + f(xj). In some cases, we followed up by assessing the pairwise concurvity between the term in question and each of the other smooth terms, i.e. x1 = f(x2), x1 = f(x3), ..., x1 = f(xj). We used the function `concurvity` in `mgcv` (Wood, 2017); the output of that function returns several measures based on the ratio of the squared Euclidean norms of vectors representing the smooth term's own space vs. the other terms. Here, we report the measure labeled *estimate* in the output of `concurvity`. Concurvity estimates range from 0 to 1, with 1 indicating that a term is fully predictable from others. A fairly common practice is to consider values greater than 0.8 to be unacceptably high (for examples of other papers on language and cognition adopting that strategy and tutorials recommending it, see e.g. Baayen & Linke, 2021; Chuang et al., 2021; Lammer et al., 2025; Saxena et al., 2022; Tomaschek & Ramscar, 2022), but we do not consider that value as a strict cutoff point; in fact, in the discussion to follow, we also scrutinize much lower concurvities.

Number of basis functions The number of basis functions constrains the degree of estimated 'wiggliness' of the prediction curves (and surfaces). We used the function `k.check` in the `mgcv` package (Wood, 2017) to check the number of basis functions.

3. Results

3.1 Type-based model

We use a published, type-based model of the data set as a baseline. The model summary, replicating Gahl and Baayen (2024), appears in Table 1, and the model's smooth terms are visualized in Figure 1. In the type-based model, predicted (average) duration increased linearly with residualized baseline duration, and in a nearly linear fashion with the proportion of prepausal tokens. The effects of NOUN BIAS and ORTHOGRAPHIC REGULARITY were non-significant. Increasing phonological neighborhood density (PND) was associated with decreasing duration in a nonlinear fashion, with the steepest predicted decrease in the lower range of PND. As for the interaction of lemma frequency and frequency ratio, the contour plot in Figure 1 shows the general gradient in the regression surface to be negative. That pattern reflects the oft-observed relationship between increasing lemma frequency and decreasing predicted duration. Increasing lemma frequency was associated with decreasing variance in duration (bottom right panel of Figure 1).

Further scrutiny of the model revealed that the deviance residuals of the model were approximately Gaussian and did not co-vary with lemma frequency or with the frequency ratio (all $|z| < 1$, all p values $> .7$), suggesting that, for all other predictors, the assumptions of normality and constant variance were met to a satisfactory degree. The two variables that did not give rise to significant effects were either not well established as predictors of spoken word duration (ORTHOGRAPHIC REGULARITY) or estimated in a manner that was likely to be highly imprecise (NOUN BIAS). In sum, the type-based model recovered a number of well established effects and was silent on variables that independently appear questionable as predictors.

The type-based model could undoubtedly be improved, but we refrain from doing so here: In the context of the present study, the model simply serves as a point of comparison of type-based vs. token-based models.

Table 1: Gaussian Location-Scale GAM fitted to the average duration of the homophone word types, retracing Gahl and Baayen (2024). te(Lemma Freq., Rel. Freq) = Tensor product of (log-transformed) target lemma frequency and ratio of (log-transformed) frequencies of the target and its homophone. AIC = -230.8.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
`Intercept` [mean]	-1.0387	0.0139	-74.7519	< .0001
`Noun Bias` =yes	0.0598	0.0169	3.5387	.0004
`Intercept` [variance]	-1.8293	0.0378	-48.4416	< .0001
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(`Proportion with Following Pauses`)	1.8335	2.2863	32.2159	< .0001
s(`Phonological Neighborhood Density`)	4.7174	5.7485	71.3616	< .0001
s(`Orthographic regularity`)	1.0000	1.0001	5.2430	.0220
te(`Lemma Freq., Rel. Freq.`)	7.2717	9.4159	108.3451	< .0001
s(`Residual Baseline Duration`)	1.1915	1.3567	122.3611	< .0001
s(`Lemma Frequency`) [variance]	2.6120	3.2966	75.3180	< .0001

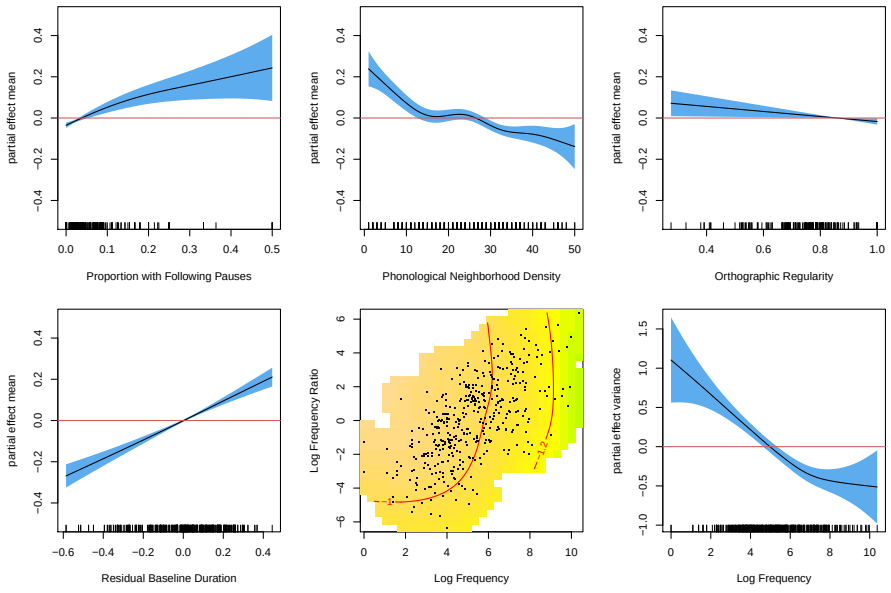

Figure 1: Partial effects according to the Gaussian Location-Scale GAM summarized in Table 1, i.e. fitted to the average duration of the homophone word types, following Gahl and Baayen (2024). Y-axis scales are fixed across panels for the partial effects on the mean, to facilitate comparison of effect sizes. Rugs indicate the unique values of predictors. In the contour plot (bottom center panel), word types, i.e. points whose coordinates reflect target lemma frequency (on the x-axis) and relative frequency (on the y-axis), are plotted as black dots.

3.2 Token-based model retracing the type-based model

We begin our discussion of token-based models with a model that is similar to the type-based model in the choice of lexical predictors, but adds token-level information. Similar to the method usually adopted in token-based analyses using Linear Mixed Effects Regression, this first token-based model includes a by-word-form random intercept. The remaining variables are identical to those in the type-based model, with the following exceptions: First, we did not include the PAUSE QUOTIENT, i.e. the proportion of tokens of each type that were followed by a pause; secondly, we included three variables coding token-specific information: the talker's sex and age, and the target's contextual probability, given the word immediately following the target word in the utterance. The summary of the model is shown in Table 2 and visualization of the regression smooths appear in Figure 2.

Table 2: GAMM fitted to the homophone tokens (n = 56,024), with a by-word-form random effect, but otherwise similar to the predictors in the type-based model (see text). Bigram prob. = Word-based bigram probability of the target, given the following word; PND = Phonological Neighborhood Density; Orthogr. = Orthographic regularity; Resid. baseline = Residualized baseline duration; Freq = target frequency; te(Freq., Freq. Ratio) = tensor product of target frequency and ratio of target and homophone frequency; pron = word form; s.1(Freq) = smooth of variance in the outcome, as predicted by frequency. AIC = 35,075.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.4137	0.0216	-65.5462	< .0001
Sex = male	-0.0608	0.0028	-21.8418	< .0001
Noun Bias = yes	0.1152	0.0089	12.9958	< .0001
(Intercept).1	-1.1408	0.0031	-369.9333	< .0001
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Bigram prob.)	2.9179	2.9945	1191.9521	< .0001
s(PND)	2.3914	2.4514	75.7883	< .0001
s(Orthogr.)	2.7376	2.8591	6.8679	.0512
s(Resid. baseline)	2.4779	2.5612	75.3640	< .0001
te(Freq., Freq.Ratio)	14.5066	15.6688	150.6620	< .0001
s(age)	2.9515	2.9982	255.2859	< .0001
s(pron)	154.2022	206.0000	4280.8009	< .0001
s.1(Freq)	3.4536	3.7788	1504.2063	< .0001

All of the predictors specific to token-based models reached significance in the expected direction, given prior corpus-based models not restricted to homophones (e.g. Bell et al., 2009; Horton et al., 2010; Yuan et al., 2006), as follows: Increasing word bigram probability was associated with shorter duration. Tokens produced by male talkers tended to be shorter than those produced by female talkers, other things being equal. Increasing talker age was associated with increasing token duration up to about age 30 and above age 55; there was no clearly discernible effect of talker age in the middle range of talker age.

With respect to the predictors that also appeared in the type-based models, the token-based model behaved similarly to the type-based model in some respects: The overall direction of the effect of PND was negative, and the effect was steepest in the lower ranges of that variable. Predicted duration increased with residualized baseline duration, as expected, although the shape of this effect was nonlinear, unlike in the type-based model. The effect of orthographic regularity was associated with a p-value exceeding our preset alpha level, as in the type-based model. Unlike in the type-based model, the effect of noun-bias reached significance, such that target forms more likely representing nouns than some other part of speech had longer predicted duration. That pattern is consistent with the idea that nouns are more often phrase-final in English than other parts of speech,

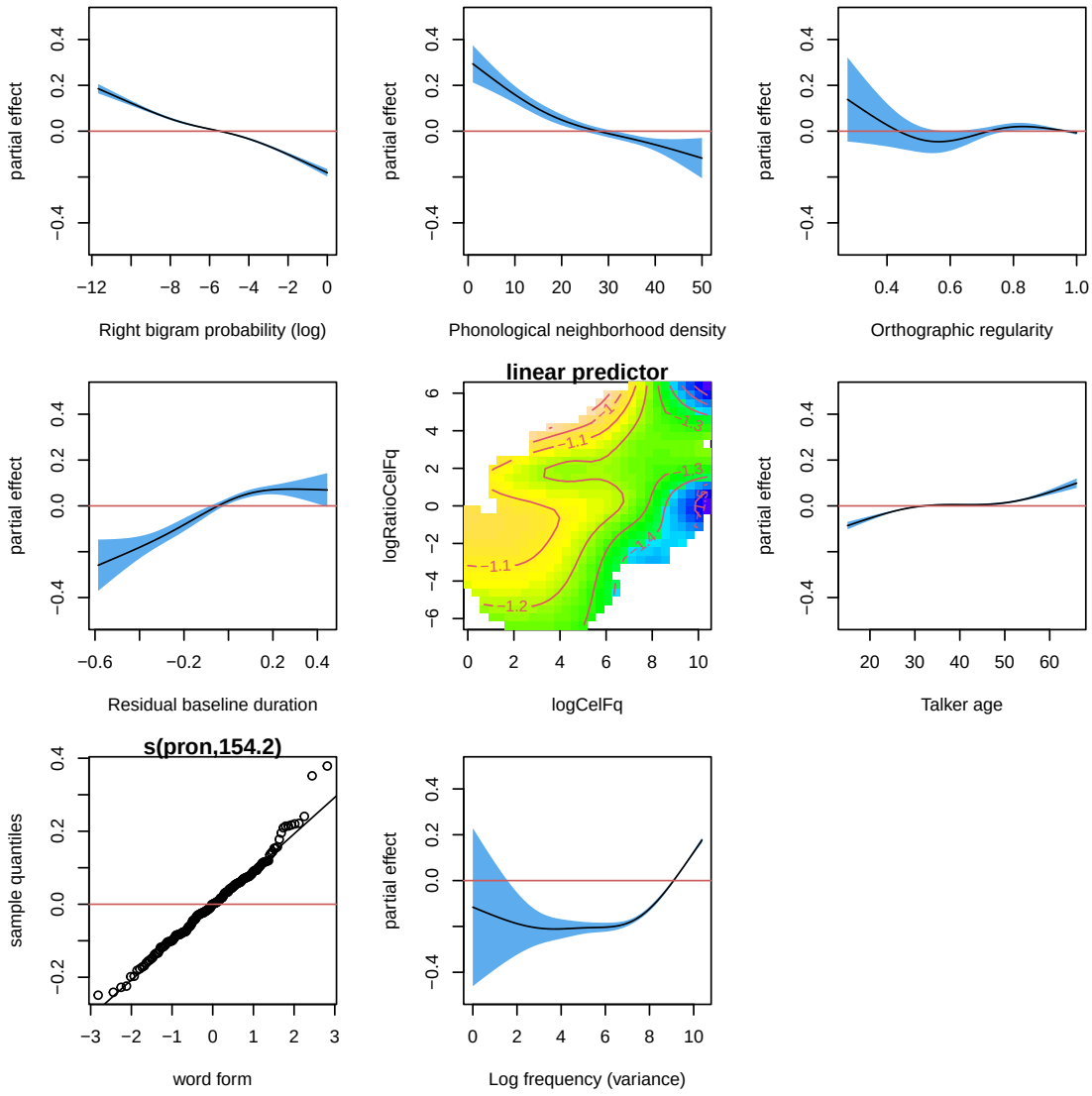

Figure 2: Partial effects according to the Gaussian Location-Scale GAMM fitted to spoken word duration of tokens summarized in Table 2.

i.e. the rationale for including this variable in the original (type-based) analysis of the homophone data (Gahl, 2008).

The token-based model also recovered the overall association of increasing frequency with shorter predicted duration, but its predictions about the interaction of target frequency with frequency ratio were more complex than in the type-based model. In the type-based model, increasing lemma frequency was associated with decreasing predicted duration across almost the entire range of the frequency ratio. In the token-based model, the effect of lemma frequency was nearly absent for words with relative frequencies of about 2, reasserting itself for words in the highest range of the frequency ratio variable. There was also some evidence for an effect of frequency ratio, such that low-frequency targets had shorter predicted durations if they had very low frequency ratios, i.e. high frequency homophone twins. To the extent that this pattern turns out to be robust, it may reflect an effect known as *frequency inheritance*: Low-frequency homophones of high-frequency words have sometimes been found to behave in some respects like high-frequency words, as if 'inheriting' the frequency of their high-frequency twins. Such 'inheritance' effects, which have been found in data based on speech errors (Dell, 1990) and latencies in semantic decisions and translation (Jescheniak & Levelt, 1994; Jescheniak et al., 2003), may coexist with lemma-specific effects of frequency on duration (cf. Gahl, 2008 for discussion).

A striking prediction of the token-based model concerns the shape of the predicted variance in duration as a function of target frequency. Recall that in the type-based model, the predicted variance decreased as target frequency increased. As seen in Figure 2, the prediction seemingly goes in the opposite direction for the token-based model. The predicted variance in duration for a given target frequency increases with target frequency. That pattern, of higher variability of high-frequency targets, is one that we take up in the discussion. The confidence region around the predicted variance is wide for low values of frequency and narrows with increasing frequency, as one would expect, given that uncertainty about an estimate (here: the estimated variance) decreases as sample size increases.

The distribution of the by-word-form random effect suggests departures from normality near the extremes of the by-word-form values. Model diagnostics further revealed that the number of basis functions was overly constrained in the case of the smooth terms for bigram probability, suggesting that the shape of the smooths as depicted in Figure 2 is underfitting the data; further exploration revealed that this issue persisted when increasing the number of basis functions to as high as 19. We return to this issue in section 3.8.1 below.

Setting aside the overly constrained estimate for the effect of bigram probability, the token-based model recovers all of the expected effects of lexical factors on word duration, and it additionally reflects plausible effects of contextual probability, talker sex, and age. The by-word-form random effect did not suggest systematic departures from normality. In sum, the token-based model looks successful at first glance.

3.3 A token-based model including additional predictors

The data set for the token-based model may well support additional predictors – and omitting such predictors may distort some of the predictions of the token-based model. Anticipating that objection, we therefore fitted a token-based model including two predictors that did not reach significance in the type-based model, *viz.* morphological complexity and biphone probability. Model summary and partial effects plots for that model are in Appendix B. Neither morphological complexity nor biphone probability reached significance based on our preset alpha level. The effect of orthographic regularity continued to be non-significant. The shape of the relationship of the remaining terms was similar to the model without the additional terms, and the predicted variance as a function of target frequency once again increased with increasing frequency. We take these results to alleviate

concerns one might have, about the omission of these variables distorting the initial token-based model (Figure 2).

3.4 The troubles begin: Concurvity

We have seen two token-based models that appear to be successful at capturing both type-level and token-level information. However, a problem with these models becomes apparent when we consider model concurvity. As noted above, concurvity is a measure of the degree to which each component of the model can be approximated by the other components. The consequences of high concurvity are analogous to those of high collinearity in linear models: Both interfere with model interpretability, by making it impossible to identify each predictor's relationship to the outcome, and by rendering individual estimates uninterpretable, i.e. liable to change substantially in response even to minor changes in modeling procedure.

Concurvity values for the type-based model (shown in Table 3) indicate high concurvity with the parametric part of the model, which suggests that combinations of the factorial predictors (i.e. talker sex and/or noun bias) are not independent of the smooths, and for the predicted variance as a function of frequency. The remaining lexical variables in the type-based model do not give rise to high concurvity.

Table 3: Concurvity of the type-based GAM in Table 1. Parm. = Parametric terms; PND = Phonological Neighborhood Density; Ortho = Orthographic regularity; 'Baseline' = Residualized baseline duration; 'Freq. tensor' = Tensor product of target frequency and frequency ratio; 'Variance' = variance predicted from frequency.

Parm.	Pause quotient	PND	Ortho	Freq. tensor	Baseline	Variance
1	.15	.31	.17	.42	.21	.97

Matters are different for the two token-based models discussed so far: As shown in Table 4, most of the terms representing type-level variables in the token-level models are highly predictable from other smooth terms in the models. The concurvity levels for PND, orthographic regularity, baseline duration, target frequency, and (in the model containing this term) biphone probability are all near .89 or even higher, rendering these terms uninterpretable. This is in stark contrast to variables whose values can vary within lexical types, i.e. local context (word-based bigram probability, for which concurvity values range from .28 to .46) and talker age (concurvity between .01 and 02, suggesting talker age to be independent of lexical variables in our sample).

Table 4 further suggests a pattern in the interplay of the by-word-form random effect with other type-level variables: When the random effect is included, PND, orthographic regularity, baseline duration, and target frequency all give rise to high concurvity. When it is not, concurvity levels for these variables go down considerably. The interplay of the by-word-form random effect with target frequency is also worth noting. Frequency is highly predictable (concurvity > .89 or higher) in all models containing the by-word-form random effect, and much less so (concurvity < .5) in models without that term. Conversely, the random effect is *least* predictable (<.4 *vs.* >.*5*) from other terms when frequency is not a predictor in the model. The by-word-form random intercept, where present, gives rise to concurvity values ranging from .37 to .57, i.e. noticeably higher than those for bigram probability and talker age, but far lower than those for variables pertaining to phonological form (PND and residualized baseline duration). The intermediate degree of concurvity of the random intercept is plausible in light of the fact that each word form (e.g. /taɪm/) in the data set receives a single posterior prediction shared by two lexical items (e.g. *time* and *thyme*) with different characteristics.

To understand whether the high model concurvity might be pinned down to specific pairs of predictors, we also inspected pairwise concurvity values. The pairwise concurvity estimates for the model in Table 2 are shown in Table 5. These estimates suggest that the by-word-form random

Table 4: Concurvity of token-based models of word token duration: 'Mirroring type-based' = 2; 'With biphone probability' = Table B1 (appendix); 'No RE' = Table C1 (appendix) 'RE, equal variance' = Table D1 (appendix); 'No RE, equal variance' = Table D2 (appendix); 'No frequency' = Table D1 (appendix); Parm. = Parametric terms; 'Bigram' = Target probability, given the previous word; PND = Phonological Neighborhood Density; Ortho = Orthographic regularity; 'Base' = Residualized baseline duration; 'Freq.' = Tensor product of target frequency and frequency ratio; 'Age' = talker age; 'RE' = by-word-form random effect; 'Var' = variance predicted from frequency; 'Biphone' = positional biphone probability.

Model	Parm.	Bigram	PND	Ortho	Base	Freq	Age	RE	Var	Biphone
Mirroring type-based (Table 2)	1	.44	1	.97	1	.97	.02	.53	1	n/a
With Biphone (Table B1)	1	.46	1	.98	1	.97	.02	.57	1	1
No RE (Table C1))	1	.30	.70	.38	.52	.48	.01	n/a	1	n/a
RE, equal variance (Table D1)	1	.42	1	.97	1	.89	.02	.51	n/a	n/a
No RE, equal variance (Table D2)	.70	.28	.64	.33	.45	.30	.01	n/a	n/a	n/a
No frequency (Table E1)	1	.38	1	.97	1	n/a	.02	.37	n/a	n/a

effect is by far the worst offender, giving rise to concurvity values at or near 1 nearly across the board; the only variables that are not fully predictable from the random effect are the two token-specific variables (bigram probability of the target word, given the next word, and talker age). The redundancy of the by-word-form random effect, and its apparent ability to render all other lexical variables in the model redundant, suggest that including that term may be a mistake. A logical response to that problem might to remove that term from the model. That is what we do next.

Table 5: Pairwise concurvity of predictors of the token-based model in Table 2. Parm. = Parametric terms; 'Bigram' = Target probability, given the previous word; PND = Phonological Neighborhood Density; Ortho = Orthographic regularity; 'Baseln' = Residualized baseline duration; 'Freq.' = Tensor product of target frequency and frequency ratio; 'Age' = talker age; 'RE' = by-word-form random effect; 'Variance' = variance predicted from frequency.

Variable	Parm.	Bigram	PND	Ortho	Base	Freq.	Age	RE	Variance
Parm.	1	.00	.00	.00	.00	.00	.00	.05	.00
Bigram	.00	1	.05	.03	.01	.04	.00	.03	.30
PND	.00	.03	1	.09	.06	.10	.00	.12	.09
Ortho	.00	.03	.18	1	.05	.06	.00	.05	.11
Base	.00	.04	.20	.15	1	.09	.00	.12	.01
Freq.	.00	.22	.56	.23	.40	1	.00	.25	1
Age	.00	.00	.00	.00	.00	.00	1	.00	.00
RE	1	.35	1	.96	1	.86	.01	1	.91
Variance	.00	.15	.07	.13	.01	.28	.00	.06	1

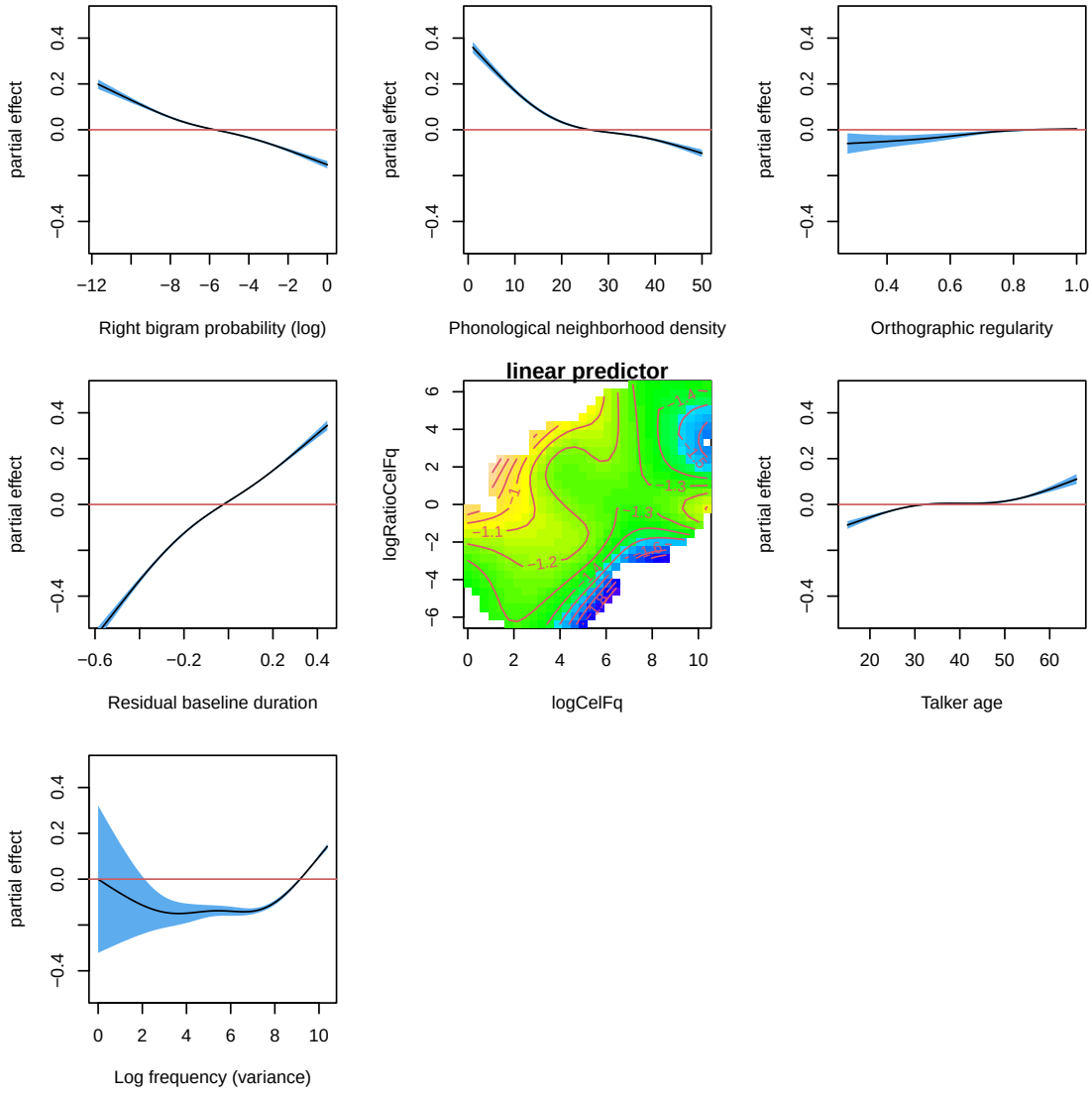

Figure 3: Partial effects according to the Gaussian Location-Scale GAM without a by-word-form random effect shown in Table C1 in the appendix.

3.5 Token-based model, without the by-word-form random effect

A token-based model without the by-word-form random effect, but otherwise identical to our initial token-based model 2, is summarized in Table C1 in Appendix C. The smooth terms are visualized in Figure 3.

The model estimates resemble those of the models with the random effect in some ways: Tokens produced by male talkers tend to be shorter than those produced by female talkers. The parametric effect of noun-bias was once again significant (unlike in the type-based model, but like in the token-based models with the random effect), in the same direction as before. The smooth term estimates were also similar to the model with the by-word-form random effects in several respects: Increasing word bigram probability and PND were associated with shorter duration. Increasing talker age was associated with increasing token duration in the lower age range (up to about age 30) and, with greater uncertainty, above age 55. Another point of similarity to the model with the random effect is the increasing predicted variance as a function of target frequency: The tensor

product of target lemma frequency and relative frequency yielded a prediction surface that was overall similar to that of the models with the random effect.

Removing the random effect also resulted in some differences, however: As one might expect, there was a noticeable narrowing of the confidence regions for the estimates of the lexical variables, i.e. PND, orthographic regularity (which reached significance in this model), and residual baseline duration. Increasing residual baseline duration, which had shown a non-linear effect (flattening in the upper ranges of the predictor) in the token-based models with the random effect, was associated with a nearly linear effect in the model without the random effect, resembling that variable's behavior in the type-based model. Removing the random effect resulted in a higher AIC (38,888.81 compared to 35,075.65 in the corresponding model with the random effect). By that criterion, the model with the random effect is preferable. As an aside, we note that the substantial change in the AIC suggests that the lexical variables, although considered well-established as predictors of lexical processing, are far from perfect as predictors of word duration in conversational speech. We do not consider this outcome to be cause for concern. The notion that the lexical variables are significant predictors at all is hardly threatened by the observation that the predictors are imperfect.

The rationale for removing the random effect was the high concurvity seemingly associated with that predictor. Did removing the random effect solve that problem? Table 4 suggests that the answer may be yes: The lexical predictors (PND, orthographic regularity, baseline duration, and frequency) were associated with much lower concurvity in the model without the random effect.

The only model component that continued to give rise to high concurvity after removing the random effect was the term predicting the variance as a function of target frequency. We therefore remove the variance estimate next, to give a random-effects model a chance to incur minimal redundancy. It is conceivable, after all, that the redundancy of the random effect was exacerbated by estimating the variance in duration as a function of frequency along with the mean. Indeed, the concurvity associated with the term predicting the variance may result from the fact that frequency appears to be predictive of duration mean and variance. This might explain why mean and variance vary in similar ways. Therefore, although the concurvity of the variance may not be the main cause for concern, a random effects model assuming equal variance might conceivably reduce concurvity to acceptable levels.

3.6 Token-based model assuming equal variance

To understand the interplay of the by-word-form random effect with estimating the variance in duration as a function of frequency, we compare models assuming equal variance with and without the random effects term. Full model summaries and partial effects plots for all predictors in the models can be found in Tables D1 and D2 and Figures D1 and D2 in Appendix D. Here, we focus on the patterns of concurvity, and on some salient differences in the smooth terms.

Table 4 shows the effect on model concurvity of including the random effect in the two models assuming equal variance: The model including the random effect incurs high concurvity for all lexical variables; as before, the only variables that are unaffected by the problem are those capturing token-level properties, i.e. word-based bigram probability of the target, and talker age. The model without the random effect shows much more acceptable levels of concurvity. In fact, in that model, even the parametric terms are not fully predictable from the rest of the model, unlike in all other models considered so far. On the 'minus' side, not modeling the variance resulted in a higher AIC (36,467.36, up from 35,075 for the otherwise identical model in Table 2), suggesting that modeling the variance is well justified from the point of view of balancing parsimony and goodness-of-fit.

How does dropping the assumption of equal variance affect the predicted relationship between the various predictors and duration? Figure 4 shows the model predictions for three variables for the two models without the variance term. For the sake of comparison, the top row of the figure repeats

the corresponding partial effects plots of the initial token-based model, i.e. the model including the by-word-form random effect and modeling the variance (Figure 2).

As Figure 4 shows, the models differ in the degree of uncertainty around those estimates, as well as in the shapes of the estimated relationships, as follows: The confidence regions around the estimates for the smooth terms are noticeably wider for the two models that include by-word-form random effects, compared to the model without that term (the bottom row in the figure). This is to be expected, given that dropping the random effect leaves it to the remaining terms to capture more of the differences across target words.

As for the shape of the relationship between predictors and predicted token duration, the effect of residualized baseline duration is non-linear in the two models with the random effect, increasing in the lower ranges of residual baseline duration and then flattening or even changing direction (though with high uncertainty) in the upper ranges. The relationship is more closely linear in the model without the random effect and without modeling the variance – resembling the linear relationship between residual baseline duration and predicted duration in the type-based model (Figure 1). The predicted effect of PND likewise changes shape: That effect is nearly linear in the model with the random effect assuming equal variance (the middle row of Figure 4, unlike in either of the other token-based models, and unlike in the type-based model. Finally, the shape of the prediction surface for the tensor product is similarly enigmatic for the models including the random effect, regardless of the inclusion of the variance term. In the model without the random effect and assuming equal variance, the prediction surface shows twin summits in the upper left quadrant (i.e. the highest values of frequency ratio in the lower ranges of target frequency) towering over the expanse of the remaining surface, which is almost entirely flat.

In sum, the high concurvity we observed for the models with the by-word-form random effect cannot be blamed on modeling the variance as a function of target frequency, as concurvity is high for both models that include the random effect. Moreover, assuming equal variance hurts the model's AIC, i.e. its goodness-of-fit balanced against parsimony.

3.7 A token-based model without frequency as a predictor

There is another possible culprit for the problems associated with the random effect, and that is the decision to include target frequency as a predictor (in the tensor product of target frequency and frequency ratio). In including frequency as a predictor, we followed common practice in corpus-based research, including some of our own. But including frequency as a predictor in token-based models might well be problematic, given that the token count itself reflects lexical frequency (albeit frequency in the corpus under analysis, rather than the CELEX frequency estimates used here). To give the random effects model another chance to account for grouping structure in the data without compromising the independence of the lexical predictors, we remove frequency from the set of predictors in the model.

The model summary and partial effects plots can be found in Table D1 in Appendix D. The pattern of significant effects is essentially unchanged in the model without frequency vs. the earlier models. The visualization of the smooths in Figure D1 suggests that the model without frequency most closely resembles the initial token-based model, i.e. the model including the by-word-form random effect. Both PND and residual baseline duration give rise to non-linear effects (generally decreasing with increasing PND, and generally increasing with increasing residual baseline duration). As before, the distribution of the by-word-form random effect suggests some departures near the extremes.

Does removing frequency as a predictor solve the problem it was intended to solve? Table 4 suggests that the answer is no: As in each previous model that included a by-word-form random effect, all lexical predictors (PND, orthographic regularity, and residual baseline duration) are nearly perfectly predictable from the model as a whole; the only non-redundant predictors are the

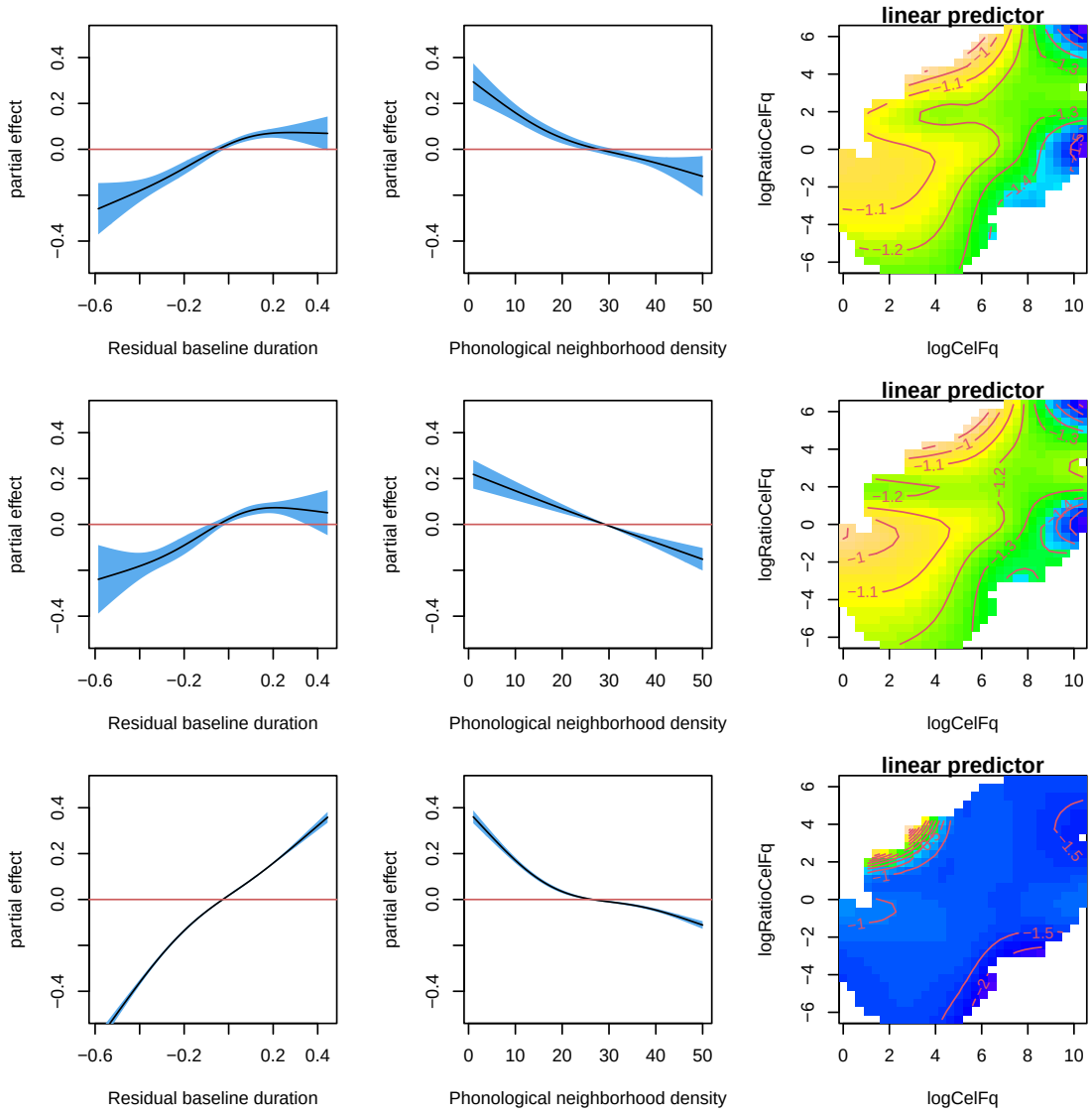

Figure 4: Partial effects of residualized baseline duration, PND, and a tensor product of target frequency and frequency ratio, according to three token-based models. Top row: A model with a by-word-form random effect and modeling variance as a function of frequency (cf. Table 2). Middle row: A model without a by-word-form random effect, also modeling variance in duration (cf. Table C1 in the appendix); Bottom row: A model without a by-word-form random effect, without modeling variance in duration (cf. Table D2 in the appendix).

token-level properties, i.e. target bigram probability and talker age. What this suggests is that, while including frequency as a predictor may be problematic, removing it does not solve the problem of high concurvity.

3.8 Model criticism

So far, we have said very little about whether the various models we considered balanced accuracy and parsimony acceptably well or satisfied the modeling assumptions. These are the questions to which we now turn.

3.8.1 Checking the number of basis functions

As mentioned above, we limited the number of basis functions for the smooth terms in our token-based models, in an effort to prevent overfitting. In our initial token-based model, we commented that the estimate of the effect of bigram probability, i.e. a token-level property, was likely over-smoothed. For the token-based model without the by-word-form random effect (cf. Table C1), `k.check` suggested that the number of basis functions was likely too low for all of the predictors, with the possible exception of talker age. Given this pattern, we increased k to 20 for all smooth terms. For the resulting model, `k.check` suggested that k was acceptably high for all predictors with the exception of bigram probability, for which k = 20 continued to be likely too low. We return to this issue in section 4.1.1 below.

Table 6: Checking the number of basis functions for the models with (left-hand columns) and without (right-hand columns) a by-word-form random intercept. RE = Random effect; s(x) = 'smooth of x'; s.1(Freq) = 'smooth of the predicted variance, given target frequency'.

	Include RE, cf. Table 2				Exclude RE, cf. Table C1			
Term	k	edf	k.index	p.value	k	edf	k.index	p.value
s(Bigram prob.)	3.00	2.92	0.90	.00	3.00	2.90	0.86	.00
s(PND)	3.00	2.39	1.00	0.41	3.00	2.98	0.96	.00
s(Orthogr.)	3.00	2.74	1.02	0.83	3.00	2.13	0.95	.00
s(Resid. baseline)	3.00	2.48	1.02	0.91	3.00	2.96	0.93	.00
te(Freq., Freq.Ratio)	24.00	14.51	1.00	0.46	24.00	18.38	0.93	.00
s(age)	3.00	2.95	1.00	0.37	3.00	2.95	0.98	.05
s(pron)	206.00	154.20						
s.1(Freq)	4.00	3.45	1.01	0.76	4.00	3.50	0.92	.00

3.8.2 Normality and homogeneity of residuals

Diagnostic plots of the models (included in the Supplementary Material) with and without by-word-form random effects suggested departures from normality in the upper ranges. In particular, residuals varied with frequency.

3.8.3 The random-effects assumption

The random-effects assumption is the assumption that the by-group adjustments (the 'random effects') are uncorrelated with the fixed effects predictors. As shown in Table 7, that assumption is violated for the token-based model in Table 2, such that all of the lexical predictors are predictive of the random intercepts. The only fixed effect whose association with the random intercept is non-significant at our alpha level is talker age.

Table 7: A model of by-word-form random intercepts in the token-based model shown in Table 2.

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	0.0134	0.0003	45.2863	< .0001
B. Smooth terms	edf	Ref.df	F-value	p-value
s(Bigram prob.)	8.7738	8.9838	67.3250	< .0001
s(PND)	8.9947	9.0000	871.9884	< .0001
s(Orthogr.)	8.9867	8.9999	161.1192	< .0001
s(Resid. baseline)	8.9962	9.0000	4536.1619	< .0001
te(Freq., Freq.Ratio)	23.9901	23.9999	2871.6014	< .0001
s(age)	3.4431	4.3121	2.1522	.0651

4. Discussion

The accessibility of corpora of naturalistic data and flexible statistical tools have fueled enormous interest in modeling linguistic experience at the token level, i.e. the level of individual utterances. Such models promise major advances over data sets of elicited speech aggregated over many utterances. 'Lab-grown' data sets restrict not only the ecological validity of conclusions, but also the kinds of questions that can be asked. The advantages of naturalistic data notwithstanding, we are sounding a cautionary note about token-based models.

We began this study by laying out four potential problems with token-based models: 'Ballot-box stuffing', model redundancy, distorted parameter estimates, and, most concerningly, assumption violations rendering model estimates valid only in a counterfactual world in which the assumptions are met. We now consider the scope of these problems. Are the issues we observed specific to the models we discussed, or are they inherent in some general property of imbalanced speech corpora, or indeed linguistic experience itself?

Assumption violations have the potential to render moot any advantages or disadvantages of token-based models: Estimates of counterfactual worlds hardly suit the kind of inquiry that make naturalistic data attractive in the first place. We therefore ask whether we can trace the observed assumption violations to any systematic property of our data. To preview our argument: There are two properties of the data that may be to blame. The first is the imbalanced nature of the sample, i.e. the fact that there are many more observations for high-frequency words than low-frequency ones. We point out some consequences of that imbalance, echoing similar observations in previous research on models with random effects. We argue that the problems extend to models without random effects. The second property of the data that may be to blame for the problems we found concerns the nature of high-frequency words. We argue that some of the problems persist in 'balanced' corpora with equal sample sizes for all items.

4.1 Four problems with token-based models

We begin by summarizing how our models fared vis-à-vis the four potential problems we identified in the Introduction.

4.1.1 Ballot-box stuffing

Do high-frequency items exert undue influence over token-based models? Interestingly, the by-token punishment for residuals did not make residuals systematically smaller for high-frequency words. In fact, as shown in the supplementary material, there is some indication that residuals actually increased with increasing target frequency. The non-independence of target frequency and model

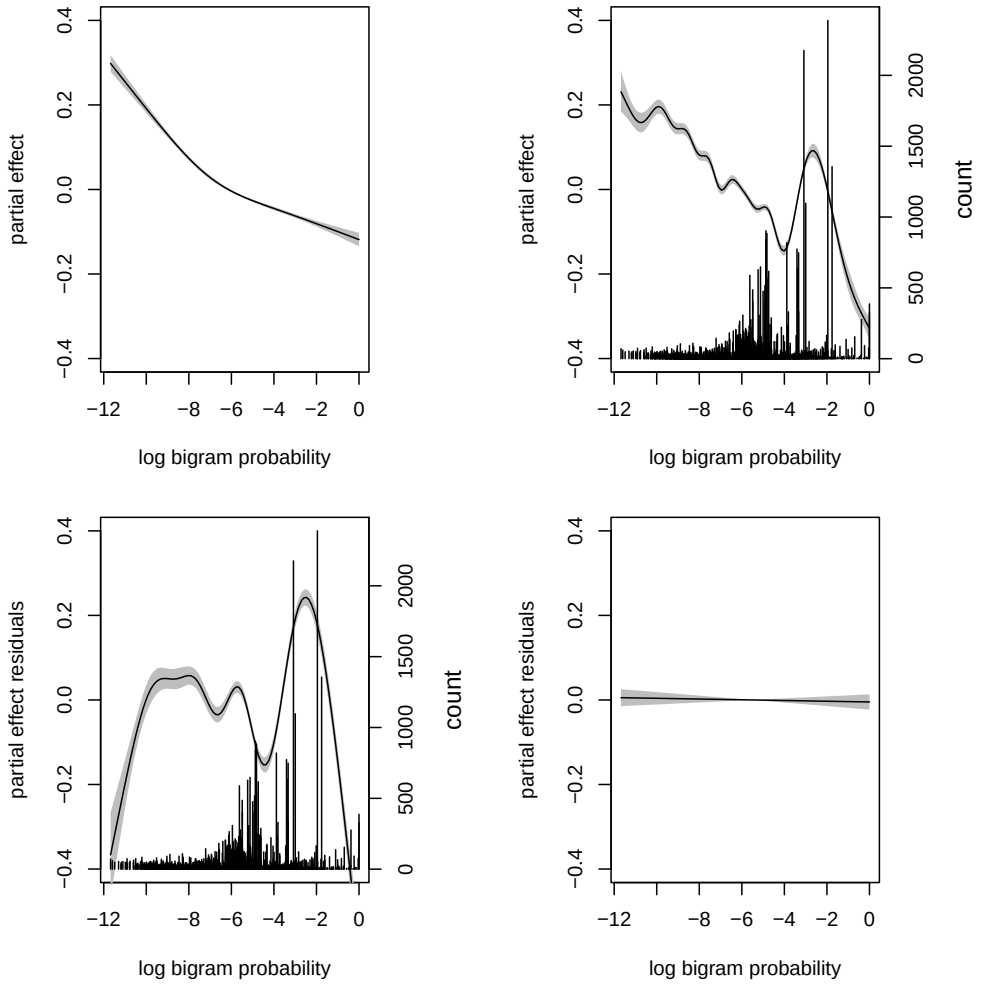

Figure 5: Smooths (upper panels) and residuals (lower panels) for $k = 4$ (left) and $k = 20$ (right).

residuals points to an assumption violation, a fact we discuss in section 4.1.4 below. Here, we focus on the issue we dubbed 'ballot-box stuffing'.

In section 3.8.1 we noted that some of the estimates in the token-based models appeared to be overly constrained in their choice of basis functions, i.e. the overall degree of wiggliness of the estimated composite function. Here, we follow up on that observation, demonstrating that this weakness of the models reflects the undue influence of high-frequency types.

In our analysis, we set k to 4 by default. With this choice of k, we allow for a moderate degree of nonlinearity. The resulting smooth for the effect of bigram probability is shown in the upper left panel of Figure 5. As mentioned earlier, the `gam.check` function of the **mgcv** package suggests that $k = 4$ is too low. Setting k to 20, the partial effect is much more wiggly (upper right panel of Figure 5), as expected. However, `gam.check` suggests that $k = 20$ is still too low. Setting k to 40 results in an even wigglier smooth (not shown), but does not change `gam.check`'s verdict that even more basis functions should be invested. Further exposing the flaws of the model where k = 4, Figure 5 (lower left panel) reveals that the residuals of that model are not equal in all ranges of (log) bigram probability, violating a modeling assumption. (Recall that, while GAMs allow the assumption of equal variances to be relaxed, this particular model only drops that assumption for one of the predictors, *viz.* target frequency.) Setting $k = 20$ removes the assumption violation: the lower right panel of Figure 5 suggests equal variance of the residuals for the entire range of (log) bigram probability. So far, everything speaks in favor of setting k to a much higher value than 4.

However, further scrutiny reveals that the values of (log) bigram probability are not evenly spread along the horizontal axis: a relatively small number of types contribute large numbers of tokens. This is illustrated by the vertical lines in the upper right and lower left panels, which are histograms visualizing the counts of observations along the range of (log) bigram probability. A mere 3.7% of values occur more than 100 times, and these jointly account for no less than 43.2% of the total number of datapoints. For example, there are 2,388 tokens for which the bigram probability is -1.97. The uneven distribution of observations creates a problem for the model: Including all tokens where the bigram probability equals -1.97 means that this particular value has strong empirical support and will strongly influence the shape of the smooth: Indeed, the smooth in the upper right panel ($k = 20$) shows a very pronounced peak just below $x = -2$. That peak is entirely absent when k is set to 4 (upper left panel). The result of increasing k is a smooth that essentially connects predictions for strongly supported predictor values. In the absence of a theory predicting high ranges of bigram probability to be associated with extremely variable outcomes (here: spoken word duration), the wiggly curve is uninterpretable and uninformative. We also note that the problem is not a problem of overfitting *per se* and is not resolved by setting the `gamma` parameter of the `gam` function to 1.4, a recommendation given by Wood (2006, p.231) as a measure that may alleviate overfitting: Setting the `gamma` parameter to 1.4 for the present data leads to a visually identical wiggly smooth (not shown here) while increasing the AIC of the model. The reason is that the issue is not so much a problem of model specification, but rather of the uneven distribution of observations.

We have discussed this particular predictor at some length. What this example illustrates more generally is that high-frequency words have the ability to influence the model by sheer force of numbers, without actually clarifying the effects of lexical predictors.

4.1.2 Model redundancy

By-item random effects have been the mechanism of choice for many researchers wishing to prevent high-frequency items from being overly influential in token-based models. As we just pointed out, this strategy appears to be unsuccessful. In addition, it appears that the by-word-form random effects exacerbate an additional problem with token-based models. We found that all of our token-based models that included the by-word-form random effect incurred high concurvity (cf. Table 4), unlike the type-based point of comparison (Table 3).

A comment on the choice of word form (e.g. /taɪm/ for *time* and *thyme*) as the grouping variable for the random effect is in order here. One might alternatively group tokens by lemma, i.e. identifying *time* and *thyme* as separate types. We opted for by-word-form random effects expressly to minimize a problem that often arises with corpus data, as pointed out in Baayen and Linke (2021): Many low-frequency lemmas are represented by so few tokens that a lemma-specific random effects structure is bound to over-fit the data. Pooling low-frequency lemmas with higher-frequency homophone

twins mitigates this problem somewhat. Relatedly, we opted for a word form-specific (rather than lemma-specific) residualized baseline duration. In Gahl and Baayen (2024), whose type-based model served as our starting point, the decision to use a word form-based residualized baseline duration measure was driven by the research question. In the current methodological context, what motivated our decision to adopt this measure was the realization that lemma-specific residualization would increase concurvity yet more.

4.1.3 Distorted parameter estimates

We commented that the token-by-token influence of high-frequency words has the potential to distort model estimates of the effects of variables other than frequency. Our comments on the wiggliness of the smooth of bigram probability in section 4.1.1 illustrate one aspect of this problem. One might limit the wiggliness of that estimate by restricting the number of basis functions, but imposing such a constraint will not elucidate the effect of contextual probability, nor will it help researchers capitalize on the richness of token-level information. An additional way in which high-frequency types can distort model estimates arises due to characteristics of high-frequency words: There are similarities among high-frequency types (see e.g. Baayen, 2011; Frauenfelder et al., 1993; Köhler, 1986; Landauer & Streeter, 1973). For example, while word-based bigram probability can in principle vary within type – for example, a given low-frequency word might be highly probable in a particular context – in actual fact, the upper ranges of bigram probability are dominated by high-frequency words. More generally, there are clusters of high-frequency word types along ranges of phonological neighborhood density, orthographic regularity, and degree of polysemy (or vagueness), to name just a few. High frequency items thus have the potential to distort estimates of the effects of these other lexical variables.

At this point, one might also ask what type of distortions or divergences across models should be considered meaningful. We believe that, ultimately, that question can only be answered in the context of specific hypotheses and research questions. Directionality changes, such as a given variable being associated with increases in the outcome in one model and with decreases in a different model, would almost certainly qualify: Theory-driven analyses often test predictions about the direction of an effect *vs.* simply checking whether a given variable significantly predicts an outcome, no matter the direction of the effect. But directionality changes are not the only type of divergence that might matter: Non-linear effects differ in 'wiggliness' as much as in directionality; the degree of wiggliness may well be central to one analysis (e.g. if one is modeling the specific trajectory of the tip of the tongue during the production of a diphthong as produced by different talkers), but irrelevant to the next (e.g. if one is interested in asking whether some variable is predictive of tongue movement at all).

The phenomenon we have dubbed 'ballot-box stuffing' has theoretical implications beyond the issue of 'wiggliness': There is a long-standing debate about the extent to which apparent effects of lexical frequency are due to frequency vs. other, correlated variables (see e.g. Gahl & Baayen, 2024; Gardner et al., 1987). Baayen (2011), for example, argues that, if other variables are allowed to do their work first, there is not much left to do for frequency. Token-based models may effectively prevent other variables from doing their work "first", because the model is obligated to minimize the residuals for the many tokens of high-frequency words.

4.1.4 Assumption violations

As mentioned in section 3.8.2 above, we found that model residuals in our token-based models increased with lexical frequency, violating the homogeneity assumption. In addition, we saw (in section 3.8.3) that the by-word-form posterior modes correlated with the fixed effects predictors. We note that the correlations of frequency with other lexical characteristics in turn make the residuals

(and the random effects in models containing those) predictable from other variables in the model, as we demonstrate in supplementary material.

Given our comments about problems traceable to the by-word-form random effect, one may ask why the residuals of the model *without* the random effect are also problematic. One reason may be that the identity of the word form is also known to the models without random effects: The model specification includes the 'residualized baseline duration', i.e. the variable reflecting the duration one would expect, given the phonemic content, setting aside what one can predict from the other lexical characteristics, such as PND. Although residualized baseline duration is a continuous variable, for the modestly-sized set of target words under discussion, it has as many distinct values as there are individual word forms, effectively identifying word forms. Removing this variable is of course possible, but would entail giving up on identifying what homophone pairs have in common, i.e. the natural experiment enabled by homophones. Indeed, removing the residualized baseline duration from the model would be appropriate if one were to take the (highly implausible) view that segmental content – e.g. the duration of a long vowel vs. a flap – was immaterial to word duration.

Some of the problems we just described have been noted before. Baayen and Linke (2021) warn against including a by-word random intercept in a model of data from the Buckeye corpus of conversational American English (Pitt et al., 2007). As is typical with such data sets, the data are highly imbalanced, containing many more tokens for high-frequency words than low-frequency ones. In fact, in Baayen and Linke (2021)'s data, nearly half of the word types occur only once. As Baayen and Linke (2021) point out, this means that there are too few observations per predictor, which in turn leads to high concurvity, and in fact, total unidentifiability, as any given observation being predicted by multiple variables. Baayen and Linke (2021) point out a second problem with the random intercepts of their model: The by-word random intercepts can be predicted from lexical variables in their data. As Baayen and Linke (2021) put it: "what should be random noise is in fact structured variation." This problem, too, is also apparent in our models, as pointed out in section 3.8.3.

4.2 Beyond homophones: Token-level models of imbalanced data

We have been utilizing a data set that has seen multiple previous analyses, in order to keep the focus on the methodological points we are making, rather than the properties of particular estimates of (type-level and token-level) variables. It might be asked whether the problems we are pointing out are specific to English homophones – or, for that matter, to the analysis of spoken word durations or other acoustic properties. We believe that this is not the case. Our points apply in principle to any analysis of lexical variables in unbalanced data sets, especially when the variables involved are correlated with type frequency. Such might be the case for analyses of reading times in running text, or lexical effects in typing speed in naturalistic samples. One example where similar issues arise is Chuang et al. (2026), in an analysis of Mandarin tones, use GAMs to predict f0 contours. In that paper, the research question and the nature of the data necessitate a token-level analysis: The research question concerns the existence of word-specific pitch contours. In order to address that question, it is first of all necessary to control for non-lexical local determinants of pitch, such as neighboring tones, position within an utterance, and syllable duration. Averages at the type level would make it impossible to control for these token-by-token properties. Secondly, Chuang et al. (2026) link token-specific pitch contours to their corresponding contextualized embeddings, i.e. another token-specific set of (semantic and form-based) properties of tokens in the context of utterances. In order to keep high-frequency words from dominating the resulting models, while ensuring that estimates of lexical (i.e. type-level) pitch signatures are based on satisfying sample sizes, Chuang et al. (2026) set an upper limit to the number of tokens per word included in the analysis, as well as a lower limit. While this solution is imperfect, for reasons we are about to discuss, it does allow lexically-specific effects to assert themselves. Importantly, the empirical

domain Chuang et al. (2026) concerns neither English, nor homophones, nor word duration, illustrating that the problems to which we are drawing attention extend beyond the particular dataset under discussion in the current paper.

4.3 Are equal samples the answer?

The problems we described are bound to arise whenever per-item samples are very unequal, even in large samples. When items have very different frequencies, the random intercepts may diverge considerably from normality (Douglas Bates, p.c.). Furthermore, when item-bound (here: by-word-form) predictors are of interest, the random intercepts are confounded with the item predictors. When predictors have values that are instantiated across multiple tokens of the same word form, whether a predictor will reach significance in the presence of a by-item random effect will depend on which values of the predictor happen to recur across these tokens. For predictors with non-repeating values, on the other hand, there is a one-to-one relation between predictor values and random intercepts. As a consequence, concurvity values will be near 1 when smoothing splines are used.

Sampling a set number of tokens for any given word type (see, for an example, Pluymaekers et al., 2005), might seem to be one strategy for addressing some or all of the problems we noted: In this way, one might keep the data from being overwhelmed by high-frequency words and remove the confound of item frequency and by-item sample size. One might take this strategy one step further and sample repeatedly, so as to avoid relying on a single sampling operation. However, repeating the sampling procedure introduces a new problem: Any given token of a high-frequency word will have a smaller chance of being drawn, and tokens of infrequent words will be included with near certainty (or complete certainty, in the case of hapax legomena). The results, over many runs, still reflect type-level properties very unevenly.

In addition, creating equal sample sizes does not remove inherent properties of high-frequency words and the confounds with other lexical properties that they give rise to. We argue that this problem cannot be overcome by using larger or more evenly sampled data sets. These are strong claims, and we suspect that analyses of additional data sets will be needed to explore these claims fully. But we see reason to believe that the properties of high-frequency words systematically affect statistical models of lexical properties, in ways that have not been fully appreciated and that may necessitate complementing token-based analyses with type-based ones. One reason for this is that lexical frequency is correlated with multiple other lexical variables (see e.g. Frauenfelder et al., 1993). Therefore, if model residuals are correlated with lexical frequency, then they will also be correlated with other lexical variables.

It might be asked why random effects, whose job it is to account for grouping structure in the data, are themselves correlated with item frequency. We think that one reason for this is that high-frequency word forms tend to be massively ambiguous (or vague), i.e. to be associated with many different meanings (and/or senses: the distinction between ambiguity and vagueness unfortunately does not help here). High-frequency word forms effectively represent many different words, as noted, for example, in Baayen et al. (2006). Along similar lines, Pimentel et al. (2020), cited in Hermalin (2025), found that more contextually predictable words tended to have more meanings/senses. As a consequence, form-level estimates for high-frequency words represent attempts to characterize wildly heterogeneous groups, and increasingly so with increasing frequency. If one believes, as we do, that word meanings and word senses matter for phonetic realization, this heterogeneity will make itself felt in the model predictions: The heterogeneity of the meanings of high-frequency forms may be partly to blame for the relationship between frequency and by-item adjustments. In this connection, it is interesting to note that the predicted variance went up as frequency increased (cf. Table 2). This might seem unexpected under the usual assumption that

uncertainty decreases with increasing sample size. But it is expected, given that high frequency forms are associated with many different meanings.

If this reasoning is correct, then drawing equal numbers of tokens (say, n = 5 per word form) again will not fix the issue: The five tokens of *hypotenuse* will still mean similar things; whereas the five tokens of *corner* will probably mean different things, having been drawn from such different contexts as *in our corner of the world*, *they backed him into a corner*, *the store is right around the corner*, *the referee awarded a corner*, and so on. The heterogeneity of high-frequency forms will still be present even if one creates equally sized samples. Careful sense disambiguation based on context-specific information about meaning may mitigate this issue. It is possible to use sense-disambiguation algorithms from Natural Language Processing to assign a discrete sense to the individual word tokens. Such a procedure does not, however, address the question of "where one sense of a word ends and the next begins" (Kilgarriff, 2006). Whether this approach solves the technical issues due semantic ambiguity and vagueness remains an open empirical question. Alternatively, contextualized embeddings can be calculated for every single token, using Large Language Models. Such contextualized embeddings estimate words' meanings in context, but the full richness of an individual speakers' communicative intentions for a given word token can only be approximated (for a comparison of these methods in a corpus-based study of Mandarin tone, see Chuang et al., 2026).

4.4 Recommendations

As stated at the outset, we are emphatically not claiming that token-level models are to be avoided under all circumstances. Setting aside the issue of assumption violations for a moment, what might be considered 'ballot-box stuffing' in one context may amount to desirable weighting of evidence in the next. We believe caution is in order particularly when the interest of an analysis concerns types, rather than tokens. Token-level models estimate token-level behavior, which can make for a mismatch between the statistical unit of analysis and the inferential target. This invites the question of how to choose between type-level and token-level models, or perhaps how to combine the two. Fundamentally, the choice of analytical target depends on one's theory of the processes that generated the data. Testing predictions about elements of a (type-based) lexicon vs. acoustic, perceptual, or situational constellations not tied to word 'types' call for different solutions (see e.g. Gahl & Baayen, 2024, for discussion). Even when the research question ultimately concerns type-level properties, token-level analyses may be a necessary element of the analysis, precisely because the process that generated the data is hypothesized to involve both type-level and token-level forces. For example, as mentioned above, the research question in Chuang et al. (2026) concerns the existence of word-specific f0 contours in Mandarin tones, a type-level property of words. In that paper, the research question and the nature of the data necessitate a token-level analysis: Averages at the type level make it impossible to control for these token-by-token properties. Superficially, the token-level analysis is necessary to control for utterance-level local determinants of pitch. More fundamentally, the authors' theory of the process that generated the data involves both type-level forces (the hypothesized lexically-specific tone contours) and utterance-level forces (such as local speaking rate and shades of meaning in specific contexts of use).

Again, the choice between analyses based on tokens and analyses based on types requires reflecting on the goal of the analysis. For example, if the goal is to predict word duration as accurately as possible with as few predictors as possible, then token-level models taking into account pre-pausal lengthening and local speech rate may be perfectly satisfactory. But these predictions will be blind to many processes that may nevertheless be operative in human language production. Token-level models can also help address certain questions about learning and development, where one may want to embrace the idea that high-frequency words are represented in the data proportionally to their frequency. During learning, word types are encountered proportional to their frequency in the

learner's experience; as a consequence, the cognitive system receives its fine-tuning predominantly from the higher-frequency words.

We conclude with a set of recommendations. Again, which of these are optimal depends on the specific (empirical and theoretical) goals of any given research project.

- If the goal is to predict properties of tokens, then, despite the problems we noted, token-level models may be the tool of choice. But prediction accuracy of these models comes at the price of disentangling contributions of individual predictors: Predictors in such models are inevitably collinear (or concurve) because high-frequency types will be represented by many replicates.
- If the goal is to evaluate models of lexical memory and retrieval of words conceived as types, then type-based models are often called for. Predictions about how many milliseconds will elapse during the production (or comprehension) of tokens is not the goal of such models, but a means to an end.
- If the goal is to understand utterance-specific factors and their interplay with lexical properties, we recommend fitting both type-based and token-based models and comparing the resulting parameter estimates and predicted values of each type of model. Consider including interactions of frequency with the other predictors in type-based models (depending on the research question) and model the variance of the outcome (Wood, 2017), keeping in mind that data sparseness near the extremes of the frequency distribution means different things in type-level vs. the token-level models: In any naturalistic corpus, there are many low-frequency word types, represented by few tokens. On the other hand, there are few high-frequency types, so there are few observations informing type-based models in the higher ranges of frequency. The consequences of aggregating data over types are difficult to identify and diagnose. Comparing type-based and token-based models can help identify those consequences.
- As a further check of the relationship between target frequency and other lexical variables, as well as in order to guard against clusters of observations (at the type level or the token level), examine the distribution of observations along the ranges of the predictors.

It is beyond the scope of the present study to evaluate these recommendations. Rather, our goal is to point out that the ability to model large sets of observations has led to a proliferation of models that may actually run counter to some of the goals of psycholinguistic research.

Type-based analyses need not forego information about contextual variables altogether. In fact, when used in combination with token-level information, they can shed light on the consequences of the accumulation of utterance-specific properties for lexical processing. For example, Seyfarth (2014) found that "Words that usually appear in predictable contexts are reduced in all contexts, even those in which they are unpredictable." Similar cumulative effects, e.g. of the positions a word most often occurs in, on its production even in other contexts have also been reported in Brown et al. (2021) and Sóskuthy and Hay (2017).

Ultimately, the decision whether to treat any variable, including frequency, as a 'lexical property' depends on hypotheses about the processes one is interested in. An analogy may be helpful here: Pine needles tend to be longer than fir needles. Even assuming the samples of needles came from an area in which fir trees were vastly more common than pine needles, or from a specific set of fir trees with more needles than a comparison set of pine trees, the difference in needle length should not be considered a 'frequency effect' on individual needles. Analogously, whether form frequency should be considered a property of tokens of, say, *thyme* depends on whether one believes that the frequency of the form [taɪm] was relevant to the processing of tokens of *thyme*.

5. Conclusion

The title of this paper alludes to Austrian Emperor Joseph II's purported characterization of Mozart's *Abduction from the Seraglio* as containing 'too many notes'. Mozart's riposte, according to anecdote, was that there were just as many notes as needed. We hope readers will consider our critical remarks to be more nuanced than the emperor's: The 'needed' number of observations may indeed equal the token count, depending on the goals of the analysis. Our conclusions have potentially broad implications. Any model based on individual tokens of items of unequal frequency is in principle subject to the issues we pointed out. These points apply to other measures besides word duration, and to other domains besides spoken word production.

6. Data Accessibility Statement

The data and analysis code for this study can be accessed at https://osf.io/7dg38/?view_only=bbb43d9a6c5e4c778b07f8905f62e56d.

Data availability statement

Goes here.

Ethics and consent

Goes here.

Acknowledgements

Goes here.

Competing interests

Goes here.

Authors' contributions

Goes here.

References

Baayen, R. H. (2011). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*, 436–461. https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512. https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding*. Benjamins. https://doi.org/10.1075/cilt.311.20baa

Baayen, R. H., & Linke, M. (2021). Generalized Additive Mixed Models. In M. Paquot & S. T. Gries (Eds.), *Practical handbook of corpus linguistics* (pp. 563–592). Springer. https://doi.org/10.1007/978-3-030-46216-1_23

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*. https://doi.org/https://hdl.handle.net/21.11116/0000-0001-91EF-E

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/bf03193014

Bates, D. M. (2005). Fitting linear mixed models in R. *R News*, *5*, 27–30. https://svn.r-project.org/R-project-web/trunk/md/doc/Rnews/Rnews_2005-1.pdf#page=27

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024. https://doi.org/10.1121/1.1534836

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers*, *19*(1), 1–9. https://doi.org/10.3758/BF03207663

Brown, E. L., Raymond, W. D., Brown, E. K., & File-Muriel, R. J. (2021). Lexically specific accumulation in memory of word and segment speech rates. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2020-0016

Chuang, Y.-Y., Bell, M. J., Tseng, Y.-H., & Baayen, R. H. (2026). Word-specific tonal realizations in Mandarin. *Language*.

Chuang, Y.-Y., Fon, J., Papakyritsis, I., & Baayen, R. H. (2021). Analyzing phonetic data with generalized additive mixed models. In M. J. Ball (Ed.), *Manual of clinical phonetics* (pp. 108–138). Routledge. https://doi.org/10.4324/9780429320903

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145. https://doi.org/10.1016/j.cognition.2017.02.001

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes*, *5*(4), 313–349. https://doi.org/10.1080/01690969008407066

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998). Resegmentation of SWITCHBOARD. *Fifth International Conference on Spoken Language Processing*.

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, *29*, 137–158. https://doi.org/10.1016/S0167-6393(99)00035-7

Frauenfelder, U. H., Baayen, R. H., & Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*(6), 781–804. https://doi.org/10.1006/jmla.1993.1039

Gahl, S. (2008). 'Time' and 'thyme' are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496. https://doi.org/10.1353/lan.0.0035

Gahl, S. (2009). Homophone duration in spontaneous speech: A mixed-effects model. *UC Berkeley Phonology Lab Technical Report*.

Gahl, S., & Baayen, R. H. (2024). *Time* and *thyme* again: Connecting English spoken word duration to models of the mental lexicon. *Language*, *100*, 623–670. https://doi.org/10.1353/lan.2024.a947037

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & Cognition*, *15*, 24–28. https://doi.org/10.3758/BF03197709

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, *1*, 517–520.

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

Hay, J. (2007). The phonetics of 'un'. *Lexical creativity, texts and contexts*, 39–57. https://doi.org/10.1075/sfsl.58.09hay

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, *139*, 83–91. https://doi.org/10.1016/j.cognition.2015.02.012

Hermalin, N. (2025). *Preliminary investigations into the communicative efficiency of logographic writing systems and written language* [Doctoral dissertation, University of California at Berkeley].

Horton, W. S., Spieler, D. H., & Shriberg, E. (2010). A corpus analysis of patterns of age-related change in conversational speech. *Psychology and Aging*, *25*(3), 708. https://doi.org/10.1037/a0019424

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(4), 824–843. https://doi.org/10.1037/0278-7393.20.4.824

Jescheniak, J. D., Meyer, A., & Levelt, W. J. M. (2003). Specific-word frequency is not all that counts in speech production. evidence from the production of homophones in dutch and german. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 432–438.

Kilbourn-Ceron, O., Clayards, M., & Wagner, M. (2020). Predictability modulates pronunciation variants through speech planning effects: A case study on coronal stop realizations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 5. https://doi.org/http://doi.org/10.5334/labphon.168

Kilgarriff, A. (2006). Word senses. In *Word sense disambiguation* (pp. 29–46). Springer.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer.

Lammer, L., Beyer, F., Riedel-Heller, S., Sacher, J., Glaesmer, H., Villringer, A., & Witte, A. V. (2025). Generalized additive mixed models to discern data-driven theoretically informed strategies for public brain, cognitive and mental health. *European Journal of Epidemiology*, 1–21.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Learning and Verbal Behavior*, *12*, 119–131. https://doi.org/10.1016/S0022-5371(73)80001-5

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38. https://doi.org/10.1017/S0140525X99001776

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*, 172–187. https://doi.org/10.1177/002383096300600306

Lohmann, A. (2018a). 'Time' and 'thyme' are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language*, *94*(2), e180–e190. https://doi.org/10.1353/lan.2018.0032

Lohmann, A. (2018b). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics*, *54*(4), 753–777. https://doi.org/10.1121/1.4987628

Pimentel, T., Maudslay, R. H., Blasi, D., & Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. *arXiv preprint arXiv:2010.02172*.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University (Distributor)*. http://www.buckeyecorpus.osu.edu

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, *118*, 2561–2569. https://doi.org/10.1121/1.2011150

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159. https://doi.org/10.1159/000090095

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Saxena, U., Mishra, S. K., Rodrigo, H., & Choudhury, M. (2022). Functional consequences of extended high frequency hearing impairment: Evidence from the speech, spatial, and qualities of hearing scale. *The Journal of the Acoustical Society of America*, *152*(5), 2946–2952.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. https://doi.org/10.1016/j.cognition.2014.06.013

Shields, L. W., & Balota, D. A. (1991). Repetition and associative context effects in speech production. *Language and Speech*, *34*, 47–55. https://doi.org/10.1177/002383099103400103

Sorensen, J. M., Cooper, W. E., & Paccia, J. M. (1978). Speech timing of grammatical categories. *Cognition*, *6*(2), 135–153. https://doi.org/10.1016/0010-0277(78)90019-7

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017. https://doi.org/10.1016/j.wocn.2020.101017

Sóskuthy, M., & Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand english. *Cognition*, *166*, 298–313. https://doi.org/10.1016/j.cognition.2017.05.032

Tanner, J., Sonderegger, M., Stuart-Smith, J., & Fruehwald, J. (2020). Toward "English" phonetics: Variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, *3*, 38. https://doi.org/10.3389/frai.2020.00038

Tomaschek, F., & Ramscar, M. (2022). Understanding the phonetic characteristics of speech under uncertainty—implications of the representation of linguistic knowledge in learning and processing. *Frontiers in Psychology*, *13*, 754395.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs [R package version 2.4].

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487. https://doi.org/10.3758/BF03195594

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic [s] in English. *Journal of Phonetics*, *11*(2), 201–206. https://doi.org/10.1016/S0095-4470(19)30816-2

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, 251–276. https://doi.org/10.1016/S0095-4470(03)00032-9

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, *90*(3), 669–692. https://doi.org/10.1353/lan.2014.0064

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), e23613. https://doi.org/10.1371/journal.pone.0023613

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wood, S. N. (2017). *Generalized Additive Models* (2nd). Chapman & Hall/CRC.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC.

Wright, R. (2004). Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology VI*, 75–87.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. *Ninth International Conference on Spoken Language Processing*.