Evaluating logistic mixed-effects models of corpus-linguistic data

Danielle Barth and **Vsevolod Kapatsinski**

University of Oregon

Abstract Evaluating the performance of mixed-effects models on the data they are trained on leads to problems in estimating model goodness. Nonetheless, mixed-effects models are preferable for corpus data, where some items have many more observations than others, because not having random effects in the model can cause fixed-effects coefficients to be overly influenced by frequent items, which are often exceptional. We explore methods for evaluating logistic mixed-effects models of both corpus and experimental data types through simulations. We suggest that the model should be tested on data it has not been trained on using some method of cross-validation and that all items (e.g., words), rather than observations, should contribute equally to estimated accuracy of the model. **Keywords:** logistic mixed-effects model, corpus, sampling, cross-validation, simulation, model comparison

1 Mixed-effects models in corpus linguistics

Linguistics is fundamentally concerned with explaining why people say what they say when they say it. Given multiple ways of conveying the same message (Labov, 1969; Sankoff, 1988), what makes speakers choose one way over another in a particular situation? The fact that there are multiple ways of conveying roughly the same intended meaning is easy to see in the case of a multilingual speaker: she might be able to say "I am a linguist" in multiple languages and would choose the language appropriate to the situation. However, it is likewise true for monolinguals. For instance, part of what causes speakers to choose the wheel of the car over the car's wheel has been shown to be the fact that the latter option results in two stressed syllables being placed next to each other (Shih et al., in press). Of course, this avoidance of stress clash is not the only predictor influencing the choice between the two genitive constructions above. Typically, many factors, semantic, syntactic, phonological, social, etc. impact the choice. This multiplicity

of influences has led linguists interested in predicting choice of expression in production, starting with Cedergren & Sankoff (1974), to turn to multiple regression models, now the main workforce of the highly related fields of quantitative corpus linguistics and quantitative / variationist sociolinguistics (e.g., Tagliamonte & Baayen, in press), both of which deal with analyses of representative databases of natural speech, or corpora.

While traditionally these regression models had only fixed-effects predictors (e.g., Tagliamonte, 2006), mixed-effects models have recently started to become the new standard (Bresnan et al., 2007; Drager & Hay, 2012; Johnson 2009, 2013). One reason (Johnson 2009, 2013), is that a valid replication of a corpus-linguistic study would not necessarily have to be a study of the same speakers, as long as the speakers come from the same speech community. Nonetheless, speakers vary around the community norm, and the effects of such variation should be taken into account, suggesting that speaker identity should be included in models of linguistic behavior and treated as a random effect.

Linguistic items are another source of variability. As Sapir (1921: 38) pointed out, "all grammars leak": no grammar is completely predictive of linguistic behavior, and part of the divergence between the grammar and the behavior comes down to the existence of exceptional linguistic items. Even if most words containing a linguistic structure behave consistently (undergo a rule affecting that structure with some constant probability), there are usually a few exceptional words that contain the structure but do not undergo the rule. This situation is typical in all languages because speakers of all languages have abundant long-term memory, allowing them to store and retrieve frequent words and phrases like *I don't know* as wholes, rather than deriving them from their parts using the grammar (e.g., Bybee, 2001). Long-term memory storage allows these frequent phrases to become exceptional in various ways. For instance, *I don't know*, unlike other *I don't Verb* phrases, can be reduced to little more than a pattern of intonation superimposed on a nasal sound (Bybee & Scheibmann, 1999).

Clark (1973) and Coleman (1964) argued persuasively that items should be treated as a random effect in psycholinguistic experiments, since the researcher samples the items from a larger population and would like to generalize to the population rather than just the sampled items. The existence of exceptions to grammatical generalizations and the fact that exceptionality seems to be concentrated in high-frequency linguistic items, provide additional reasons for treating items as a random effect in studying grammar. However, corpus data differs from psycholinguistic data in two ways, which raise additional questions. First, in psycholinguistic data, every item is presented to every subject the same number of times. In corpus data, the items that are more frequent in the language will be observed more frequently. Second, using an item frequently has been argued to lead to articulatory reduction and semantic bleaching (Bybee, 2002, 2003; Gafos & Kirov, 2009; Pierrehumbert, 2001; Schuchardt, 1885). To the extent that this is true, high-frequency items are not just more likely to deviate from the sample mean than low-frequency items; they are also likely to deviate from the mean in a dif-

ferent direction than low-frequency items. Thus in a corpus study, the sample is biased to sample items that are likely to be exceptional. We describe these challenges in more detail in the following section. The rest of the paper is devoted to addressing the implications of these challenges for the use of mixed-effects models in corpus linguistics by means of Monte Carlo simulations.

2 The challenges of corpus data

As discussed above, one of the main challenges of corpus data is that the data are not nicely balanced: some speakers might contribute huge portions of the database, while others may contribute only one or two observations; in any corpus, there are a few extremely frequent words, while most words occur only once (Baayen, 2001). If we want corpus data to be representative of people's linguistic experience (Biber, 1993; McEnery et al., 2006: 19), the lack of balance across speakers is unavoidable: we all have a few family members and close friends who account for most of our linguistic input.

Lack of balance across linguistic items is likewise unavoidable: linguistic expressions are subject to a rich-get-richer effect. The more often a word is used, the more likely it is to be re-used in the future: frequent words come to mind more readily than infrequent words in language production (as demonstrated by Oldfield & Wingfield, 1965, in a picture naming task). As we would expect from a variable subject to rich-get-richer positive feedback loops (Yule, 1925; Simon, 1955; Merton, 1968; Barabási & Albert, 1999), word frequencies display a power-law distribution (Zipf, 1949; Baayen, 2001; Newman, 2005).

As Bresnan et al. (2007), Gerard et al. (2010), and Sonderegger (in press), among others, note, mixed-effects models are in fact hoped to address this issue more effectively than fixed-effects-only models.

> "An elegant solution to this sort of situation, where the data come in groups of very different sizes, is offered by mixed-effects logistic regression, via what Gelman and Hill (2007:252–9) call "partial pooling." In Model 2, in addition to fixed effects of frequency and structure, we will include a random effect of Prefix on the intercept. This means we assume that the intercept... —the intrinsic likelihood of change, in log-odds—differs by prefix class... Roughly speaking, the intercepts for small classes are inferred based on the proportion of changed words in large classes." (Sonderegger, in press)

There is widespread agreement that the identity of a particular word is an important predictor in the study of grammar and language change: word frequency does not account for all differences among words (e.g., Pirrrehumbert 2002). Perhaps, the least controversial example is that verbs differ in their preferences for various syntactic constructions (Briscoe & Copestake, 1999; Goldberg, 1995;

Pinker, 1984). Importantly, frequency interacts with our ability to learn that a certain item is exceptional: the problem of data sparsity that partial pooling is intended to solve is not just a problem for linguists analyzing a corpus. It is also a problem for language learners.

Stefanowitsch (2008) notes that we are much more confident about the intransitivity of the frequent verb *disappear* than about the intransitivity of the infrequent *vanish* (**He disappeared it.* is judged as being less grammatical than **He vanished it.*; see also Boyd & Goldberg, 2011 for supportive experimental data). Frequency influences our ability to learn that an item is exceptional (Bybee, 2002; Erker & Guy, 2012), making the between-item differences more pronounced among high-frequency words. For instance, Erker & Guy (2012) show that some high-frequency verbs in Spanish favor the omission of subject pronouns while others disfavor it, whereas all low-frequency verbs are alike. Raymond & Brown (2012) show that fricative reduction is exceptionally productive in some high-frequency words (ones that tend to occur in casual speech) and exceptionally unproductive in others (words that tend to occur in more formal registers). These results are very much in line with partial pooling: the lowest-frequency words cannot be exceptional because language learners cannot estimate individual probabilities (of co-occurring with some construction or undergoing some change) for those verbs and must instead rely on the lexicon-wide probabilities, probabilities that are based on data pooled across individual words.

As the quotation from Sonderegger (in press) above makes clear, an assumption underlying treating item as a random effect is that items for which we have little data should behave like items for which we have much data. However, this assumption is not necessarily true of linguistic items because frequent linguistic items are seen to undergo changes that infrequent ones do not undergo. In particular, they are likely to be articulatorily reduced (e.g., losing their final [t]'s in English, Bybee, 2002). The effect of frequency on reduction may make partial pooling problematic for studying reductive grammatical processes unless frequency is taken into account. Partial pooling assumes that low-frequency words should behave like high-frequency words. Lexical diffusion theory suggests that low-frequency words should not behave like high-frequency words. Not only are the data unbalanced but the items for which we have much data may not behave like the items for which the data are sparse. Thus even the simplest model of grammatical behavior should include item frequency, and not just item identity, as a predictor. This special status of frequency is why we focus on this predictor in the present paper.

3 Current approaches to model evaluation in corpus linguistics

To the extent that a corpus is a representative sample of the recorded speakers' productions, the within-speaker predictors of a regression model of that corpus, along with the associated coefficients, can be thought of as a description of the speakers' production grammar, thus evaluation and comparison of alternative regression models is fundamental to the linguistic enterprise (Cedergren & Sankoff, 1974; Labov, 1969). Model evaluation is concerned with two related questions: 1) how much room is there for improvement over the current model, i.e., should we look for additional predictors or have we described the grammar fully?, and 2) if we found an additional predictor, should we go for the more complex model with that predictor, or the simpler model that lacks it?

There are essentially three approaches to model evaluation that are widely used in the corpus-linguistic literature employing mixed-effects models, with some studies combining multiple methods (e.g., Bresnan & Ford, 2010; Riordan, 2007). The first method, and one we advocate against, measures how well the model fits the data (Baayen, 2008: 281; Keune et al., 2005; Kothari, 2007; Lohmann, 2011; Theijssen, 2009) using measures of accuracy like Somers' Dxy and C (index of concordance). Here models are evaluated and, often implicitly, compared on how often the value they predict matches the value observed. Importantly, the models are tested on the same data they are fit to. This is what we argue against (along with Pitt & Myung, 2002, among others). We show that the accuracy of a mixed-effects model is unaffected by scrambling the values of a predictor that we know to be useful for predicting the values of the dependent variable: when the fixed-effects predictor does not do a good job, a random-effects predictor can rise to the occasion and capture the same variance.

We are not the first to make this observation in corpus linguistics. Antić (2010, 2012) and Yao (2011) have likewise noticed this anecdotally and proposed that the goodness of a fixed-effects predictor is verified if it can capture some of the variance that would otherwise be attributed to a mixed-effects predictor. In other words, a more complex model can be accepted over a less complex model if the more complex model uses fixed effects to capture some of the variance that the simpler model attributes to random effects. For example, Antić (2010:42) found that detecting a prefix was easier, with respect to reaction time, when the word containing the prefix could be easily decomposed into morphemes. She compared two mixed-effects models, one containing measures of compositionality, her theoretically motivated fixed effects, and one that did not contain them. Adding the measures of compositionality to the model did not make the model have a better fit to the data but the measures accounted for 88% of the variance that the model without these measures attributed to the random effect of item. She concluded that compositionality does have an effect on prefix detection reaction times, a conclusion that depends on the random effects "stepping in" to capture the residual vari-

ance when the fixed effects are not there to account for it. We show that a random-effects predictor does indeed capture more variance when the values of a useful fixed-effects predictor are scrambled. However, this raises the question of whether there is any reason to include the random effects in the model: on the Antić/Yao approach, a fixed-effects predictor is accepted if it covaries with the dependent variable whether or not random effects can capture the same variance.

Following Pitt & Myung (2002), among others, we argue for making decisions about which predictors to include in a model by testing the model on unseen data, a practice attested but uncommon in the corpus literature (Bresnan & Ford, 2010; Riordan, 2007; and Theijssen et al., in press). We fit the model to data containing a random subset of the levels of a random-effects predictor and test on the rest of the data. In this case, the model can only do a good job predicting the values of the dependent variable in the test data by means of fixed effects: the levels of the random effects in the test data are unfamiliar from training. On this measure, again, a mixed-effects model that contains a meaningful fixed-effects predictor performs better than a model that does not, indicating that random-effects are still useful for estimating the coefficients associated with the fixed-effects predictors. The model tested on unseen data has only coefficients associated with fixed effects. However, we show that for highly unbalanced datasets typical of corpus data, the coefficient estimates are much better (in that they are more predictive) when the random effect is included in training the model. These results reaffirm the usefulness of mixed-effects modeling for corpus data (Johnson 2009, 2013) as well as the usefulness of testing models on data they were not fit to.

4 Simulations

4.1 Against fit to training data

We created one thousand replications of a simple corpus study, in which there is one fixed-effects predictor and one random-effects predictor potentially influencing the probability of reduction. The random-effects predictor was the identity of the word, while the fixed-effects predictor was word frequency. In every replication, both predictors had a real influence. The effect of word is shown in Figure 1: some words are associated with one value of the dependent variable while others are associated with the other. The distribution of probabilities is bimodal, as it should be, given the frequent observation that speakers may be uncertain about the grammatical behavior of unknown words and yet be certain about the behavior of the words they know (e.g., Albright & Hayes, 2003; Kapatsinski, 2010a). Nonetheless, the distribution of the corresponding logits (log-odds) is quasi-normal,

making logistic regression appropriate for analyzing the data: if we compare the distribution for every replication to a normal distribution with the same number of items, mean and standard deviation using the Kolmogorov-Smirnov test, it comes out significant ($p<.05$) on only 2.9% of the time in the sample of 1000 replications. Two normal distributions with the same mean, standard deviation and number of observations come out as significantly different 3.2% of the time in the sample. Thus, the distribution of item effects is quasi-normal in logit space.

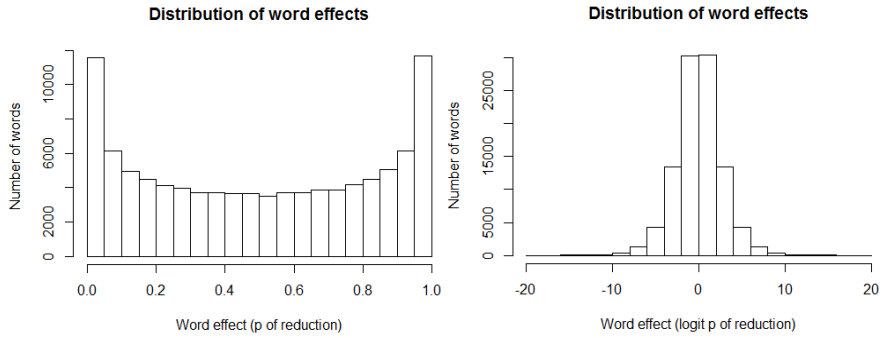

Fig. 1. Distribution of word effects in terms of probability of choosing the 'reduced' value of the dependent variable (bimodal) and the corresponding logits (normal) for the words in the simulations.

For every replication, we ran two mixed-effects models: one model included the real fixed-effects predictor while the other included its randomly scrambled version. Models were fit using the lme4 package in R (Bates et al. 2012). By comparing the two models to each other, we can then determine the predictive power of the fixed-effects predictor, in other words, how much variance in the data is explained by frequency. Figure 2 shows that the manipulation worked: the fixed effect of frequency was greater than the fixed effect of scrambled frequency: the coefficients for frequency center around 1.5 while those for scrambled frequency center on zero.

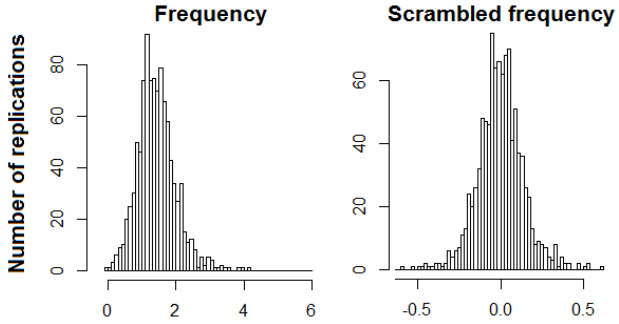

Fig. 2. The distribution of the effect of frequency (left) vs. scrambled frequency (right).

For every one of the 1000 replications, two versions of the dataset were created. In one version, the number of observations for each word was directly proportional to its frequency: frequent words contributed more observations. Frequency (and hence number of observations) was distributed according to the Pareto distribution (Newman, 2005), illustrated in Figure 3. This version is more likely to resemble the samples obtained in a corpus study, where more observations are found of frequent items (e.g., Bresnan et al., 2007). In the other version, each word contributed the same number of observations, which was equal to the mean number of observations contributed by a word in the other condition. (Thus, mean number of observations per word was equated in the two sampling conditions). The balanced sampling scheme is usual in experimental studies.

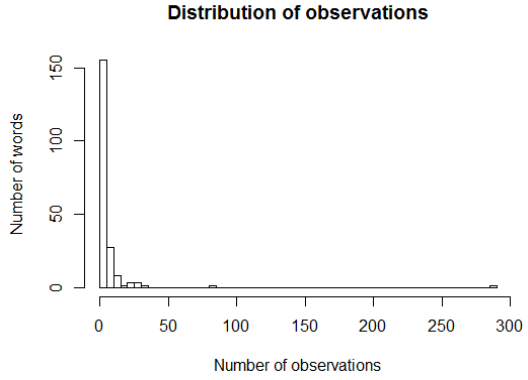

Fig. 3. The distribution of number of observations across words for a corpus-like sample, in which number of observations for a word is proportional to its frequency.

The first way we compared the models was using the Concordance index on the entire dataset to which the models were fit (as suggested in Baayen, 2008). The distributions of concordance indices for balanced sampling and skewed sampling are shown in Figure 4. As Figure 4 shows, the distributions are virtually identical: scrambling the frequency predictor did not decrease the index of concordance of the model despite reducing the coefficient associated with the fixed-effects predictor to zero. This result supports the hypotheses of Antić (2010, 2012) and Yao (2011) that a real predictor can nonetheless fail to contribute to how well a mixed-effects model fits the data. The random-effects predictor steps in to capture the variance that the scrambled fixed-effects predictor is no longer capturing. The result holds for both corpus data, where number of observations is correlated with values of the fixed-effects predictor in question, frequency, and experimental data where the two are uncorrelated and the number of observations per cell in the design is controlled. We believe that this result conclusively argues against using fit to training data to evaluate mixed-effects models.

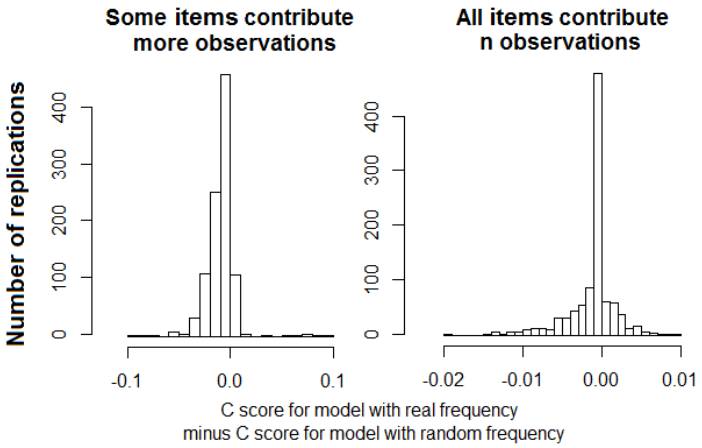

Fig. 4. Histograms of differences in C scores between a model that has real frequency values and a model that contains scrambled frequency values as a predictor. The C scores for the model containing the real predictor are no greater than the C scores of the control model that contains scrambled frequency values as a predictor.

4.2 Testing the models on withheld data.

4.2.1 Leave-one-out cross-validation (LOOCV), lexical frequency and lexical diffusion.

We consequently turned to comparing the models on predicting the data they have not been fit to by means of leave-one-out cross-validation (LOOCV). On every replication, we cycled through all the items training the model on all items but one and testing it on the remaining, withheld, item. As before, we compared models containing real frequency values to models containing scrambled ones. Scrambling was done between items. In other words, in both models with original frequency values and models with scrambled frequency values, frequency was constant across observations of a single linguistic item. Thus the hierarchical structure of the dataset was preserved.

We tried two ways of assessing model accuracy: 1) standard LOOCV, where we simply computed the proportion of observations a model got right, or 2) averaging accuracies computed within items. With the first way, each observation gets a vote in testing the predictive power of the model; with the second way, each item gets a vote. On both theoretical and empirical grounds, we suggest that the latter way of assessing models is superior for models containing grammatical pre-

dictors and frequency. First, let us consider frequency as a predictor. Consider what happens when a high-frequency word is withheld from training and used for testing the model. Given that frequency distributions are like that shown in Figure 3, i.e., there are few high-frequency words and many low-frequency words (Zipf, 1949), eliminating a high-frequency word from training is likely to leave no other word of similar frequency to use in estimating the frequency effect. By contrast, eliminating a low-frequency word from training has virtually no effect on estimating the frequency effect: there are plenty of other words with the same frequency to use. As a result, accuracy of a model containing an effect of frequency should be lower when the model is tested on high-frequency words than when it is tested on low-frequency words. Yet, with corpus-like sampling, a high-frequency word contributes many more observations to the sample than does a low-frequency word. Thus, the effect of frequency is likely to be seriously underestimated by this procedure. Averaging accuracy values computed within items allows each item to contribute equally to model evaluation, alleviating this problem.

The empirical difference between these two ways of carrying out LOOCV for frequency is shown in Figure 5, where the x-axis shows how well the random-effects model performs and the y-axis shows how well the fixed-effects model performs. Especially in the case of the GLM, giving a vote to each observation reduces differences between the predictiveness of real frequency and the predictiveness of scrambled frequency and reduces certainty about the magnitude of the frequency effect.

Consider now a grammatical predictor. There is much evidence that high-frequency items are especially likely to be exceptional. This is, of course, expected if high-frequency items can be retrieved from the lexicon directly rather than derived through the grammar (e.g., Bybee, 2001; Kapatsinski, 2010a, 2010b), or if language learners are able to estimate item-specific random slopes of grammatical predictors for high-frequency items (Erker & Guy, 2012). Testing a model of the grammar on high-frequency items it has not been trained on is then likely to underestimate the power of the model, and allowing a single high-frequency item to contribute more tokens to the test set than a single low-frequency item would exacerbate the issue. As argued by Berko (1958), grammatical models may best be tested on unseen items that have a frequency of zero in the language since these are least likely to be retrieved from the lexicon directly. However, this kind of testing is obviously impossible with corpus data: an item that has a frequency of zero will not be seen in a corpus of naturalistic speech. We compromise by making each word contribute to the estimated accuracy equally. Since there are few high-frequency words (Zipf, 1949), they do not greatly influence the estimated accuracy if one word gets one vote as opposed to each observation getting a vote.

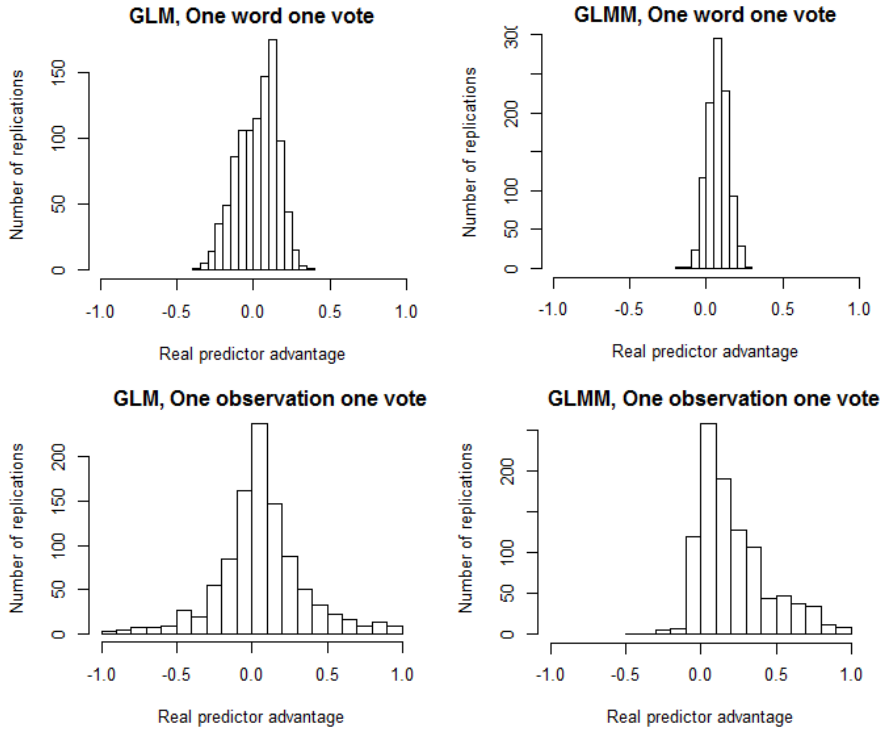

Fig. 5. The effect of frequency depending on type of model (fixed-effects only GLM vs. mixed-effect GLMM) and type of evaluation. Corpus-like sampling assumed throughout.

4.2.2 Mixed-effects vs. fixed-effects and sampling.

Since the test item was withheld from training, the models' random-effects predictor 'item' does not have a level corresponding to the test item. Thus, for every mixed-effects model, we extracted the intercept and the coefficient associated with the fixed effect of frequency and generated predicted values for the unseen item using those coefficients and the observed frequency value of the unseen item.

The model we are testing now on unseen data is a fixed-effects-only model: the only predictor is frequency, which is a fixed effect. However, the coefficient for that predictor could be estimated using the mixed-effects model. We reasoned that the resulting estimate may be more accurate than one that would be obtained by fitting a fixed-effects-only model to the training data because it would partial away variability due to individual items. To test this idea, we also generated a fixed-effects-only near-equivalent to every mixed-effects model using the glm() function in R. The fixed-effects-only model had no random-effects predictors.

Figure 6 shows how well the resulting models fit the test data compared to models that contained scrambled frequency values. The data suggest that having the random effect in the model that is fit to the training data makes the coefficient associated with the fixed effect more accurate when the number of observations varies, as in corpus studies, rather than being controlled, as in experimental work. The GLM and GLMM perform very similarly for the balanced sample: the right (unscrambled) predictor is preferred by both models; model predictions across replications correlate at r =.95, and both models prefer the right predictor approximately 90-91% of the time ($\chi^2(1)$= 0.28, p=.6). However, it should be noted that the fixed-effects-only model prefers the right model more strongly, i.e., the difference in predictive power between the model containing real frequency values and the model containing scrambled frequency values is greater in the fixed-effects-only model than in the mixed-effects model (55% of the time, which is significantly above 50%, $\chi^2(1)$= 10, p<.002). Thus for balanced sampling, the two types of models display similar predictive power; and when both models prefer the real predictor, the fixed-effects-only model prefers it more strongly. Since the scrambled predictor is equally non-predictive in both models, this result suggests that the coefficient estimate for the real predictor is (slightly) better without the random effect in the model.

However, the mixed-effects GLMM is much more likely to prefer the model containing the real frequency values when number of observations varies along with frequency: 87% for GLMM vs.62% for GLM ($\chi^2(1)$= 154.3, p<.00001). We should note that corpus-like sampling reduced the likelihood of detecting a real effect of frequency even for the mixed-effects model (91% vs. 87%; $\chi^2(1)$=7.59 , p=.006) but the mixed-effects model shows a much better ability to cope with corpus-like sampling. These data suggest that fixed-effects coefficient estimates are best estimated using mixed-effects models for corpus data even though they need to be extracted and tested on data that the model was not trained on.

What is it then about the corpus-like data that causes fixed-effect coefficients estimated on the basis of mixed-effects models to be so superior to those estimated by the GLM? In particular, is it crucial that number of observations is correlated with the values of the fixed-effects predictor in question? Is it crucial that predicted values are biased in favor of the value of the dependent variable associated with values of the fixed-effects predictor for which we have more observations? Is it crucial that distinct observations of an item always have the same value on the fixed-effects predictor? Or is the fact that different items contribute different numbers of observations sufficient to make the coefficient estimates based on the mixed-effects model superior? We argue that unbalanced sampling across levels of the random-effects predictor is sufficient, hence the superiority of coefficient estimates based on mixed-effects models should be true for all kinds of predictors in corpus studies.

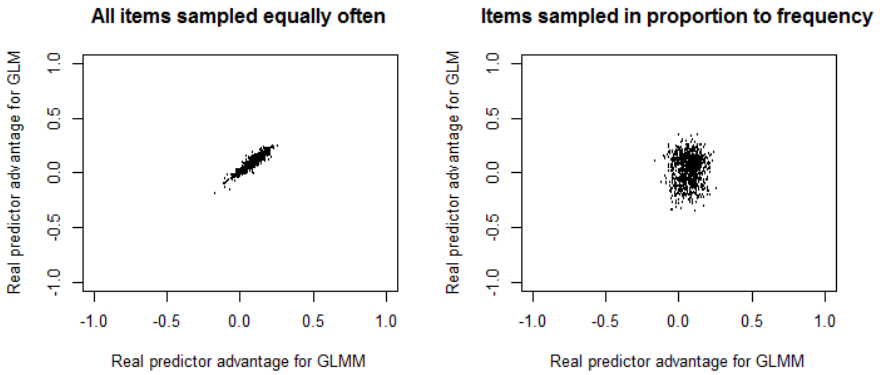

Fig. 6. Mixed-effects (GLMM) vs. fixed-effect-only (GLM) models and sampling. Negative values indicate that the model (inappropriately) prefers having the scrambled frequency values while positive ones indicate that it prefers real frequency values. With a balanced (experiment-like) dataset to be trained on, GLM and GLMM perform similarly; with corpus-like sampling, the GLMM is superior in distinguishing the real predictor from the scrambled predictor.

To address these questions, we switched the actual fixed-effects predictor to be uncorrelated with frequency and item by randomly selecting the level of the predictor (either "1" or "0") for every observation. The dependent variable was then a function of both item identity and the value of the predictor. In this new simulation, the fixed-effects predictor differed from the fixed-effects predictor in previous simulations in all potentially relevant respects: 1) it was binary rather than continuous, 2) the value of the predictor was not correlated with number of observations of that values, 3) it was also not predictable based on values of the random-effects predictor, and 4) because of this, the dependent variable was equally likely to take on either value, and was predicted to be equally likely to do so. Nonetheless, the same results were obtained (Figure 7): for a balanced sample, fixed-effects coefficients derived from both models are very similar (on the left in Figure 7, r =.88), while for the corpus-like sample the mixed-effects model is much more likely to be successful in preferring the real fixed-effects predictor over the scrambled version (the mixed-effects model prefers the real predictor 87% of the time, while the fixed-effects-only model prefers it only 21% of the time; the difference between the models is highly significant: $\chi^2(1)=554.5$, $p<.00001$). The fixed-effects-only model frequently has no preference for the real predictor (62% of replications). Corpus-like sampling thus greatly diminishes the predictive power of the fixed-effects-only model (reducing number of correct predictor selections from 99.5% to 21%). Corpus-like sampling does also hurt the mixed-effects model (99.5% vs. 87% correct predictor selections, $\chi^2(1)=5.2$, $p=.02$) but to a much lower extent. Thus we suggest that the mixed-effects model is highly preferred for unbalanced corpus-like samples whatever the predictor type.

There is a strong correlation in real predictor advantages between the two models when the sampling scheme is balanced but the correlation breaks down when sampling is corpus-like. We suspect that this is due to the fixed-effects-only model often basing its coefficient estimates to a large extent on the large proportion of data that come from the one or two highly frequent items in the sample, which sometimes works out (when those items are typical) but often does not, whereas the mixed-effects model is able to partial out the variance due to items.

To summarize, a number of results were obtained here with two very different types of predictors. First, with balanced sampling, there is no advantage to incorporating a random effect of item even if there is such an effect when the aim is to generalize to new items. Second, with unbalanced, corpus-like sampling, random effects are essential for obtaining fixed-effects coefficient estimates that can be used to predict behavior on unseen items. In additional simulations, which we cannot report here due to space constraints, we have also verified that these results hold for a continuous predictor that is uncorrelated with frequency of an item and can vary within an item, for a binary predictor that cannot vary within an item and does not correlate with frequency, and for binary and continuous predictors that have effects whose slope varies across items.

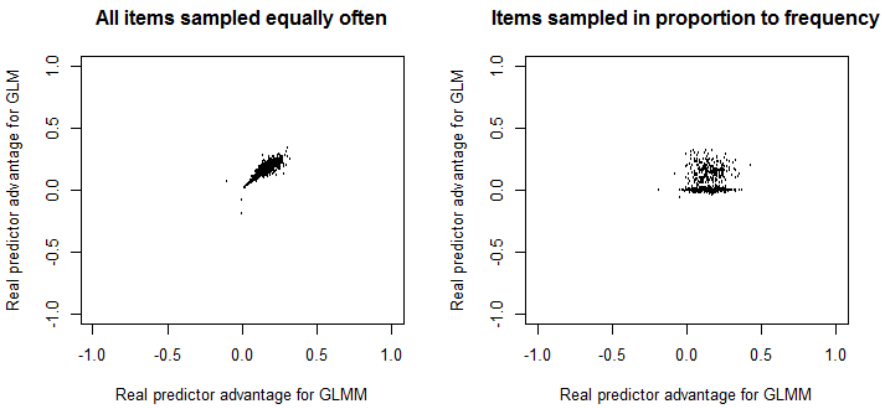

Fig. 7. Mixed-effects (GLMM) vs. fixed-effect-only (GLM) models and sampling. For balanced samples, both models work very similarly and always prefer the real predictor; for corpus-like samples, the fixed-effects-only model (GLM) is much more likely to fail to prefer the real predictor whereas the mixed-effects model (GLMM) continues to almost always prefer the actual predictor.

5 Discussion

A possible limitation of the present work is the exclusive use of leave-one-out cross-validation. We do not have a strong commitment to leaving only one of the

items out at any one time. Leaving out, say, a tenth of all the items is a plausible alternative (Kohavi, 1995) that should be explored in future work. The bootstrap (Efron & Tishirani, 1993), which differs from cross-validation in that the data are sampled with replacement in constructing test and training sets, is another possibility that we have been reluctant to pursue as it would mean that the model would occasionally be tested on data it has been trained on. Finally, it may also be possible to improve on the method of Yao (2011) and Antić (2010, 2012) by developing an inferential method for comparing models on how much of the variance captured by a random effect is reduced by adding a fixed effect.

However, given the goal of selecting or evaluating a model of *grammar* that is likely to have generated the corpus, we believe in the importance of allowing each word/item to vote equally and testing the model on novel items that it was not trained on. The reason is that we believe that much of the time known words and expressions are retrieved from memory rather than generated via the grammar (Albright & Hayes, 2003; Bybee, 2001; Kapatsinski, 2010a, 2010b). We are particularly persuaded by the findings that a grammatical generalization may lose productivity, i.e., stop being extended to novel items, while having few if any exceptions in the lexicon (Becker et al., 2011; Kapatsinski, 2010a, 2010b). This is only possible if the existing words instantiating the generalization are retrieved from memory directly, allowing the generalization to remain true of the language's lexicon or corpus while not being picked up on by native speakers of the language.

Given the ubiquity of full-form retrieval, the grammar, or the set of *productive* generalizations should be most obvious in low-frequency words (Bybee, 2001). Unfortunately, we do not know *how* low in frequency a word should be to not be retrieved from the lexicon directly, thus we cannot non-arbitrarily exclude words from the test set, on which a model of grammar is evaluated, or allocate different words different numbers of votes as a function of frequency. However, given that word frequency distributions are Zipfian (Baayen, 2001; Newman, 2005; Zipf, 1949), with the majority of words in a corpus being relatively infrequent words, allocating each word one vote, as we have done here, makes it really important for a model of grammar to predict how speakers produce low-frequency words. Of course, a corpus study should ideally be supplemented by an experiment with truly novel words, since a corpus may not even contain words that are infrequent enough not to be retrieved directly from the lexicon (Berko, 1958).

An additional argument for allocating votes to words, rather than observations, is that many studies found that the productivity of a generalization is influenced by the number of distinct words instantiating that generalization, rather than the number of instances of such words ("type" rather than "token" frequency, e.g., Albright & Hayes, 2003; Bybee, 2001; Richtsmeier, 2011). For instance, additional exposures to a word do not make other words that contain similar sound sequences more acceptable (Richtsmeier, 2011). These findings suggest that evaluating a grammar by the number of word *types* it gets right (without being trained on them) may also have some psychological reality.

6 Conclusions

Corpora are valuable data. In them, researchers find real use of language. However, as argued in section 2, the 'realness' of this data type presents unique challenges in model evaluation. Here, we have argued that training and testing of models of this data should be done on different sections of the dataset, using cross-validation techniques, and that prediction accuracy of models should be evaluated based on item types, rather than tokens. We have also argued that unbalanced, corpus-like datasets demand the use of mixed-effects models, even if the model is tested on levels of the random effect it was not trained on.

In our first simulation, we have shown that evaluating mixed-effects models on fit to the training data does not allow one to select a model containing a real fixed-effects predictor over a model that contains a predictor whose values have been randomly scrambled. This result provides evidence against comparing mixed-effects models on how well they can fit the data: if the fixed-effect predictors in the model are unpredictive, the random effects can "step in" to capture the variance, allowing the model to still fit the data well. Nonetheless, this model would be useless in generalizing to unseen items that it was not trained on.

When we test a model on unseen items, the random-effect of item may a priori appear to be of no use. Our second simulation shows that this intuition is incorrect. Even when models are tested on how well they predict behavior on unseen items, a mixed-effects model containing item identity as a predictor still has higher predictive power than a fixed-effects-only model. This advantage of mixed-effects models is seen for highly unbalanced samples typical of corpus data but not for balanced experimental designs, confirming that incorporating a random-effects predictor is particularly important when the number of observations across the values of that predictor is unbalanced. Without the random effect of item, the model may base its estimate of how the speaker will behave largely on high-frequency items. This is a particularly important issue with linguistic data, since high-frequency linguistic items are precisely the ones that are likely to be exceptional. Mixed-effects models provide a way to deal with this issue by partialling out the variance due to individual items, allowing for better generalization to unseen items.

References:

Albright, A., & B. Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90, 119-161.

Antić, E. 2012. Relative frequency effects in Russian morphology. In *Frequency effects in language learning and processing*, ed. by S. Th Gries & D. Divjak, 83-108. Berlin: Mouton de Gruyter.

Antić, E. 2010. The representation of morphemes in the Russian lexicon. UC Berkeley Dissertation.

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R. H., D. J. Davidson, & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language*, 59, 390-412.
Barabási, A.-L. & R. Albert. 1999. Emergence of scaling in random networks. *Science*, 286, 509–512.
Bates, Dougles, Martin Maechler, & Bin Dai. 2012. lme4: Linear mixed-effects models using S4 classes. R package version .999999-0.
Becker, M., N. Ketrez, & A. Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 87, 84-125.
Berko, J. 1958. The child's learning of English morphology. *Word*, 14, 150-77.
Biber, D. 1993. Representativeness in corpus design. *Literary & Linguistic Computing*, 8, 243-57.
Boyd, J. K., & A. E. Goldberg. 2011. Learning what not to say: the role of statistical preemption and categorization in "a"-adjective production. *Language*, 87, 1-29.
Bresnan, J., A. Cueni, T. Nikitina, & R. H. Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. by G. Bouma, I. Kraemer, & J. Zwarts, 69-94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.
Bresnan, J., & M. Ford. 2010. Predicting syntax: Processing Dative constructions in American and Australian varieties of English. *Language*, 86, 168-213.
Briscoe, T., & A. Copestake. 1999. Lexical rules in constraint-based grammars. *Computational Linguistics*, 25, 487-526.
Bybee, J. 2003. Mechanisms of change in grammaticization: the role of frequency. In *Handbook of historical linguistics*, ed. by B. Joseph & R. Janda, 602-23.Oxford: Blackwell.
Bybee, J. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation & Change*, 14, 261-90.
Bybee, J. 2001. *Phonology and language use*. Cambridge, UK: Cambridge University Press.
Bybee, J., & J. Scheibmann. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in American English. *Linguistics*, 37, 575-96.
Cedergren, H., & D. Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language*, 50, 333-55.
Clark. H. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, 12, 335-59.
Coleman, E. B. 1964. Generalizing to a language population. *Psychological Reports*, 14, 219-26.
Drager, K., & J. Hay. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation & Change*, 24, 59-78.
Efron, B. & R. Tibshirani. 1993. *An introduction to the bootstrap*. London: Chapman & Hall.
Erker, D., & G. R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, 88, 526-57.
Gafos, A., & C. Kirov. 2009. A dynamical model of change in phonological representations: The case of lenition. In *Approaches to phonological complexity*, ed. by F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupé, 219-40. Berlin, New York: Mouton de Gruyter.
Gelman, A., & J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
Gerard, J., F. Keller, & T. Palpanas. 2010. Corpus evidence for age effects on priming in child language. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 218-23.
Goldberg, A. E. 1995. *Constructions*. Chicago: Chicago University Press.
Johnson, D. E. 2013. Progress in regression: Why sociolinguistic data calls for mixed-effects models. Ms.
Johnson, D. E. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language & Linguistics Compass*, 3, 359–83.
Kapatsinski, V. 2010a. Rethinking rule reliability: Why an exceptionless rule can fail. *Chicago Linguistic Society*, 44(2), 277-291.

Kapatsinski, V. 2010b. What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. *Corpus Linguistics & Linguistic Theory*, 6, 157-215.

Keune, K., M. Ernestus, R. Van Hout, & R. H. Baayen. 2005 Social, geographical, and register variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics & Linguistic Theory*, 1, 183-223.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, *International Joint Conference on Artificial Intelligence (IJCAI)*, *Volume 2*, 1136–1143. Morgan Kaufmann.

Kothari, A. 2007. Accented pronouns and unusual antecedents: A corpus study. *Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue*, 150-57. Antwerp: Association for Computational Linguistics.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language*, 45, 715-62.

Lohmann, Arne. 2011. Help vs. help to: A multifactorial, mixed-effects account of infinitive marker omission. *English Language & Linguistics*, 15, 499-521.

McEnery, T., R. Xiao, & Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. London, New York: Routledge.

Merton, R. K. 1968. The Matthew effect in science. *Science*, 159, 56-63.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46 5, 323–351.

Oldfield, R. C., & Wingfield, A. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 4, 272-81.

Pierrehumbert, J. 2002. Word-specific phonetics. In *Laboratory Phonology 7*, ed. by C. Gussenhoven & N. Warner, 101-40. Berlin: Mouton de Gruyter.

Pierrehumbert, J. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In *Frequency and the emergence of linguistic structure*, ed. by J. Bybee & P. Hopper. Amsterdam: John Benjamins.

Pinker, S. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pitt, M. A., & I. J. Myung. 2002. When a good fit can be bad. *Trends in Cognitive Sciences*, 6, 421-25.

Raymond, W. D., & E. L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In *Frequency effects in language learning and processing*, ed. by S. Th Gries & D. Divjak, 35-52. Berlin: Mouton de Gruyter.

Richtsmeier, P. 2011. Adults learn phonotactic probabilities from word-types not word-tokens. *Journal of Laboratory Phonology*, 2, 157-84.

Riordan, B. 2007. There's two ways to say it: Modeling nonprestige *there's*. *Corpus Linguistics & Linguistic Theory*, 3, 233-79.

Sankoff, D. 1988. Sociolinguistics and syntactic variation. In *Linguistics: The Cambridge Survey, Vol. IV*, ed. by F. Newmeyer, 140-61. Cambridge, UK: Cambridge University Press.

Sapir, E. 1921. *Language*. New York: Harcourt Brace & Co.

Schuchardt, H. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.

Shih, S., J. Grafmiller, R. Futrell, & J. Bresnan. In press. Rhythm's role in genitive construction choice in spoken English. In *Rhythm in phonetics, grammar and cognition*, ed. by R. Vogel & R. van de Vijver, Berlin: Mouton de Gruyter.

Simon, H. A. 1955. On a class of skew distribution functions. *Biometrika*, 42, 425-40.

Sonderegger, Morgan. In press. Testing for frequency and structural effects in an English stress shift. *Berkeley Linguistics Society*, 36.

Stefanowitsch, A. 2008. Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19, 513-531.

Tagliamonte, S. 2006. *Analyzing sociolinguistic variation*. Cambridge, UK: Cambridge University Press.

Tagliamonte, S., & R. H. Baayen. In press. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation & Change*.

Theijssen, D. 2009. Variable selection in logistic regression: The British English dative alternation. In *Interfaces: Explorations in logic, language and computation*, ed. by T. Icard & R. Muskens, 87-101. Berlin, Heidelberg: Springer.

Theijssen, D., ten Bosch, L., Boves, L., Cranen, B., & van Halteren, H. in press. Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics & Linguistic Theory*.

Yao, Y. 2011. The effects of phonological neighborhoods on pronunciation variation in conversational speech. Ph.D. Dissertation, UC Berkeley.

Yule, G. U. 1925. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society B*, 213, 21-87.

Zipf, G. K. 1949. *Human behaviour and the principle of least effort*. Reading, MA: Addison-Wesley.