Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure

Emily Morgan (eimorgan@ucsd.edu)

Roger Levy (rlevy@ucsd.edu)
Department of Linguistics, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0108 USA

Abstract

Most models of choice in language focus on broadly applicable generative knowledge, treating item-specific variation as noise. Focusing on word order preferences in *binomial expressions* (e.g. *bread and butter*), we find meaning in the item-specific variation: more frequent expressions have more polarized (i.e. frozen) preferences. Of many models considered, only one that takes expression frequency into account can predict the language-wide distribution of preference strengths seen in corpus data. Our results support a gradient trade-off in language processing between generative knowledge and item-specific knowledge as a function of frequency.

Keywords: Bayesian modeling; binomial expression; frequency; word order

Introduction

A pervasive question in language processing research is how we reconcile generative knowledge with idiosyncratic properties of specific lexical items. In many cases, the generative knowledge is the primary object of study, while item-specific idiosyncrasies are treated as noise. For instance, in modeling the dative alternation, Bresnan, Cueni, Nikitina, and Baayen (2007) take care to demonstrate that effects of animacy, givenness, etc. on structure choice hold even after accounting for biases of individual verbs. But the verb biases themselves are not subject to any serious investigation. Here we present evidence that patterns within the item-specific variation are meaningful, and that by modeling this variation, we not only obtain better models of the phenomenon of interest, we also learn more about language structure and its cognitive representation.

Specifically, we will develop a model of word order preferences for *binomial expressions* of the form *X and Y* (i.e. *bread and butter* preferred over *butter and bread*). Binomial ordering preferences are in part determined by generative knowledge of violable constraints which reference the semantic, phonological, and lexical properties of the constituent words (e.g. short-before-long; Cooper & Ross, 1975; McDonald, Bock, & Kelly, 1993), but speakers also have idiosyncratic preferences for known expressions (Morgan & Levy, 2015; Siyanova-Chanturia, Conklin, & van Heuven, 2011). Binomial expressions are a useful test case for modeling idiosyncracies because their frequencies can be robustly estimated from the Google Books n-grams corpus (Lin et al., 2012). Here we will demonstrate that explicitly modeling these expressions' idiosyncrasies both produces a better predictive model for novel expressions and also constrains our theory of these expressions' cognitive representations.

Specifically, we identify two reasons why such a model is advantageous:

1. Models identify both rules and exceptions.

One intrinsic reason that modeling idiosyncrasies is advantageous is because identifying exceptions can help identify rules. In a traditional linguistic setting (e.g. identifying rules for past tense formation), we rely upon intuition to determine what is the grammatical rule and which verbs should be treated as exceptions. In the case of binomial expressions, we likewise expect there to be exceptions to the rules, particularly for frequent expressions. For example, there is in general a strong constraint to put men before women; however, *ladies and gentlemen* is preferred over the reverse due to its conventionalized formal use. But compared with past tense formation, the rules that determine binomial ordering are far more complex and gradient, such that using traditional linguistic analysis to determine the full set of rules is not viable. In this case, we require our model not only to identify what the rules are but simultaneously to determine which expressions must be treated as exceptions. Having such a model is useful for empirical cognitive science, e.g. for disentangling the effects of people's generative knowledge from effects of their item-specific linguistic experience on language processing (Morgan & Levy, 2015).

2. Models relate cognitive representations to language-wide structure.

As a further benefit, models can help us understand how structural properties of the language relate to people's cognitive linguistic representations. In particular, let us look at the distribution of preferences for binomial expressions taken from a subset of the Google Books corpus (described later in Creating the Corpus.) Each binomial can be assigned a preference strength corresponding to how frequently it appears in alphabetical order, from 0 (always in non-alphabetical order) to 0.5 (perfectly balanced) to 1 (always alphabetical). Binomials which always or nearly always appear in one order are said to be *frozen*. The distribution of preference strengths is shown in Figure 1. Preferences have a multimodal distribution with modes at the extremes as well as around 0.5. This distribution poses a challenge to standard models of binomial preferences. As we will show later, standard models predict only a single mode around 0.5. In other words, the true distribution of binomial expressions includes more frozen binomials than standard models predict. As we develop a model that accounts for this multimodal distribution, we will see that this language-structural fact puts constraints on our theories of individuals' cognitive representations of binomial expressions.

In the remainder of this paper, we first describe how we developed a new corpus of binomial expressions. We then explore a variety of models with differing levels of ability to model item-specific idiosyncrasies. Finally, we return to the issue of how these models inform us about cognitive representations of language.

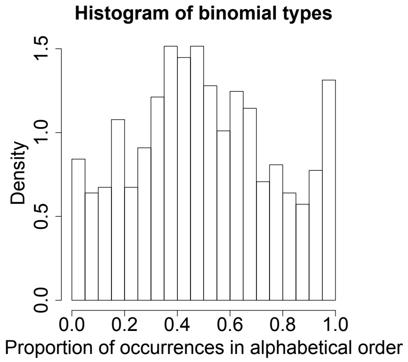


Figure 1: Binomial preferences are multimodally distributed in corpus data

Creating the Corpus

We extracted all *Noun-and-Noun* binomials from the parsed section of the Brown corpus (Marcus, Santorini, Marcinkiewicz, & Taylor, 1999) using the following Tregex (Levy & Galen, 2006) search pattern:

```
/^N/=top < (/^NN/ !$, (/,/ > =top) .
((CC <: and > =top) . (/^NN/ > =top)))
```

This pattern finds all Noun-*and*-Noun sequences dominated by a Noun Phrase which are not preceded by a comma (to exclude the final pair in lists of more than two elements), a total of 1280 tokens.

Binomials were coded for a variety of constraints, originally described by Benor and Levy (2006) but restricted to the subset determined to be most relevant for predicting ordering preferences by Morgan and Levy (2015):

Length The shorter word (in syllables) comes first, e.g. *abused and neglected*.

No final stress The final syllable of the second word should not be stressed, e.g. *abused and neglected*.

Lapse Avoid unstressed syllables in a row, e.g. *FARMS and HAY-fields vs HAY-fields and FARMS*

Frequency The more frequent word comes first, e.g. *bride and groom*.

Formal markedness The word with more general meaning or broader distribution comes first, e.g. *boards and two-by-fours*.

Perceptual markedness Elements that are more closely connected to the speaker come first. This constraint encompasses Cooper and Ross's (1975) 'Me First' constraint and includes numerous subconstraints, e.g.: animates precede inanimates; concrete words precede abstract words; e.g. *deer and trees*.

Power The more powerful or culturally prioritized word comes first, e.g. *clergymen and parishioners*.

Iconic/scalar sequencing Elements that exist in sequence should be ordered in sequence, e.g. *achieved and maintained*.

Cultural Centrality The more culturally central or common element should come first, e.g. *oranges and grapefruits*.

Intensity The element with more intensity appears first, e.g. *war and peace*.

The metrical constraints, Length and No final stress, were automatically extracted from the CMU Pronouncing Dictionary (2014), augmented by manual annotations when necessary. Word frequency was taken from the Google Books corpus, counting occurrences from 1900 or later. Semantic constraints were hand coded by two independent coders (drawing from the first author and two trained research assistants). Discrepancies were resolved through discussion.

For each binomial, we obtained the number of occurrences in both possible orders in the Google Books corpus from 1900 or later. Items containing proper names, those with errors in the given parses, those whose order was directly affected by the local context (e.g. one element had been mentioned previously), and those with less than 1000 total occurrences across both orders were excluded from analysis, leaving 594 binomial expression types.

Models

We will develop four models of binomial ordering preferences: a standard logistic regression, a mixed-effects logistic regression, and two hierarchical Bayesian beta-binomial models. All are based on the idea of using logistic regression to combine the constraints described above in a weighted fashion to produce an initial preference estimate for each binomial. The models differ in whether and how they explicitly model the fact that true preferences will be distributed idiosyncratically around these estimates. The standard logistic regression includes no explicit representation of item-specific idiosyncrasies. The mixed-effect logistic regression includes random intercepts which account for item-specific idiosyncrasies, but which are constrained to be distributed normally around the initial prediction. The two Bayesian models assume that item-specific preferences are drawn from a beta distribution whose mean is determined by the initial prediction. In the first of these models, the concentration of the beta distribution is fixed, while in the second, it varies with the frequency of the binomial in question.

Evaluation

One obvious criterion for evaluating a model is how well it predicts known binomial preferences (i.e. the corpus data). For this, we report $R^2(X,\hat{X})$ as well as mean L1 error, $\frac{1}{N}\Sigma_{i=1}^{N}|\hat{x}_i - x_i|$, where $\hat{x}_i$ is the model prediction for how often binomial i occurs in a given order, and x_i is the true corpus proportion.

In addition to considering model predictions for each individual item, we want to consider the overall distribution of preferences within the language. As we will see, a model can provide good predictions for individual items without correctly capturing the language-wide multimodal distribution of these expressions' preference strengths. Thus our second desideratum will be the shape of the histogram of expression preferences.

Logistic regression

Logistic regression is the standard for modeling syntactic alternations, both for binomial expressions specifically (e.g. Benor & Levy, 2006; Morgan & Levy, 2015) as well as other syntactic alternations (e.g. Bresnan et al., 2007; Jaeger, 2010). Thus we begin by constructing a baseline logistic regression model. Benor and Levy have argued that one should train such a model on binomial types rather than binomial tokens because otherwise a large number of tokens for a small number of overrepresented types can skew the results. While agreeing with this logic, we note that to train only a single instance of each type is to ignore a vast amount of data about the gradient nature of binomial preferences. As a compromise, we instead train a model on binomial tokens, using token counts from the Google Books corpus, with each token weighted in inverse proportion to how many tokens there are for that binomial type, i.e. a type with 1000 tokens will have each token weighted at 1/1000. In this way, we preserve the gradient information about ordering preferences (via the diversity of outcomes among tokens) while still weighting each type equally. The constraints described above are used as predictors. Outcomes are coded as whether or not the binomial token is in alphabetical order.

For this and all future models, predictions are generated for all training items using 20-fold cross validation. Results for all models can be seen in Figure 2. While the logistic regression model does a reasonable job of predicting preferences for individual items, it does not capture the multimodal distribution of preference strengths seen in the corpus data. We proceed to consider models in which item-specific idiosyncrasies are modeled explicitly.

Mixed-effects regression

By far the most common method in language modeling for accounting for item-specific idiosyncrasies is mixed-effects regression models (Jaeger, 2008). Formally, this model assumes that idiosyncratic preferences are distributed normally (in logit space) around the point estimate given by the fixed-effects components of the regression model.

We train a mixed-effect logistic regression on binomial tokens using the `lme4` package in R. We use as predictors the same fixed effects as before, plus a random intercept for binomial types. As described above, the fitted model now predicts a *distribution*, rather than a single point estimate, for a novel binomial. To make predictions for our (cross-validated) novel data, we sampled 1000 times from this distribution for each item. The histogram in Figure 2(c) shows the full sample distribution across all items. In order to generate point estimate predictions for computing L1 and R^2 (shown in Figure 2(b)), we take the sample median for each item, which optimizes the L1 error.

Including random intercepts improves neither our point estimates nor our language-wide distribution prediction. Apparently, the normal distribution of the random intercepts is not well suited to capturing the true distribution of binomial preferences. In particular, for a given item, the normality of random effects in *logit* space leads to predictions that are skewed towards the extremities of *probability* space.[1]

Hierarchical Bayesian beta-binomial model

Having seen that normally distributed random intercepts do not adequately capture the distribution of item-specific preferences, we introduce the beta distribution as a potentially better way to model this distribution. The beta distribution, defined on the interval $[0,1]$, has two parameters: one which determines the mean of the draws from the distribution, and one which determines the *concentration*, i.e. whether draws are likely to be clustered around the mean versus distributed towards 0 and 1. For example, for a beta distribution with a mean of 0.7, a high concentration implies that most draws will be close to 0.7, while a low concentration implies that roughly 70% of draws will be close to 1 and 30% of draws will be close to 0. When we treat the output of the beta distribution as a predicted binomial preference, a high concentration corresponds to a pressure to maintain variation while a low concentration corresponds to a pressure to regularize.

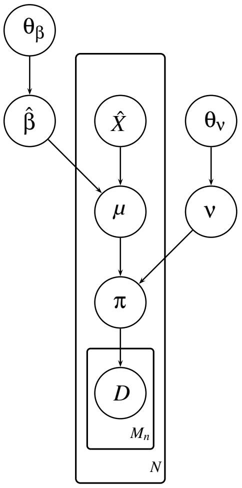
In order to incorporate the beta distribution into our model of binomial preferences, we combine the logistic regression and the beta distribution in a hierarchical Bayesian model (Gelman et al., 2013), as shown in Figure 3. For each item, the model determines a mean μ via standard logistic regression, using the same predictors as before. The model also fits a concentration parameter ν. These two parameters determine a beta distribution from which the binomial preference π is draw. Observed data is drawn from a binomial distribution with parameter π.

We fit this model using the `rjags` package in `R` (Plummer, 2003). After a burn-in period of 2000 iterations, we run for 2000 more iterations sampling every 20 iterations. In order to predict novel data, we fix the point estimates for the regression coefficients $\hat{\beta}$ and the concentration parameter ν. We then sample 1000 draws of π for each item. As with the mixed-effects model, the histogram in Figure 2(c) shows the full sample distribution, while point estimates (the sample median) are used to calculate L1 error and R^2 (Figure 2(b)).

This model performs better on L1 and R^2 than the mixed-effects model, but still worse than the initial logistic regression. The predicted histogram shows hints of the multimodal distribution seen in corpus data, but is overall too flat.

---

[1] An alternative method of prediction for novel items would be to take the median random intercept in logit space, i.e. to set all random intercepts to 0. This method yields results that are very similar to—but all-around slightly worse than—the original regression model.



N: # unordered binomial types
M_n: Frequency of binomial n
$\hat{X}$: Predictors (i.e. generative constraints)
$\hat{\beta}$: Regression coefficients
θ: Uninformative priors
$\mu = \dfrac{1}{1+e^{-\hat{X}\cdot\hat{\beta}}}$
$\nu \sim exp(\theta_\nu)$
$\pi \sim \mathrm{Beta}(\mu,\nu)$
$D \sim \mathrm{Binomial}(\pi, M_n)$

Figure 3: Our initial hierarchical Bayesian beta-binomial model. The set of nodes culiminating in μ implements a standard logistic regression. The output of this regression determines the mean of the beta distribution (with ν determining the concentration) from which π and finally the observed data itself is drawn.

Beta-binomial with a variable concentration parameter

A crucial fact that we have not taken into account in previous models is the role of frequent reuse in shaping expressions' preferences. In particular, the degree to which an expression takes on a polarized preference may depend upon its frequency. We build upon the beta-binomial model in the previous section by parameterizing the concentration parameter by the frequency of the (unordered) binomial expression:

$$\nu = exp(\alpha + \beta \cdot \log(M_n)) \tag{1}$$

where M_n is the total number of occurrences of binomial n in both orders. Training and testing of the model are identical to above.

We find that $\beta = -0.26$ is significantly different from 0 ($t_{99} = -94; p < 2.2 \times 10^{-16}$), indicating that the concentration parameter changes significantly as a function of frequency: less frequent expressions have more dense distributions while more frequent expressions have more polarized distributions, as shown in Figure 5. We find that this model generates the best predictions of all our models, producing a marginally significant improvement in both L1 ($t_{593} = 1.86; p = 0.06$) and R^2 (by fold $t_{19} = 1.76; p = 0.09$) relative to the initial logistic regression. Moreover, it correctly predicts the multimodal distribution of expression preferences.

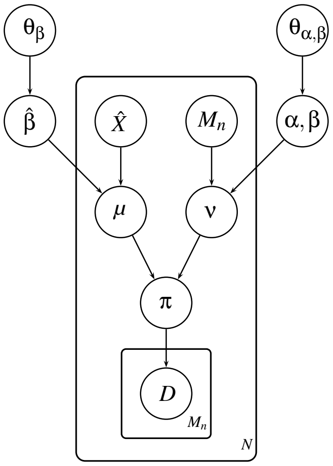


Figure 4: Hierarchical Bayesian beta-binomial model with variable concentration parameter

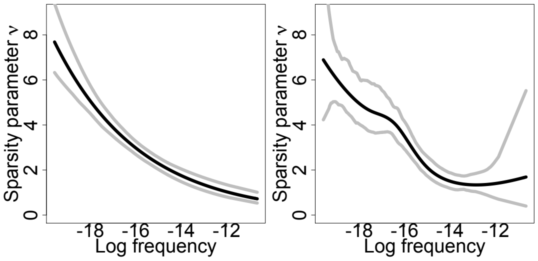


Figure 5: Concentration parameter ν as a function of frequency with 95% confidence intervals. (Left) Parameterization given in Eq. 1. (Right) Alternate parameterization with cubic splines, for comparison.

Discussion

Overall, we found that all models made approximately similarly good best-guess predictions for binomials they weren't trained on, but the frequency-sensitive beta-binomial model was clearly superior in predicting the language-wide distribution of idiosyncratic binomial-specific ordering preferences. This model also indicates that more frequent binomials are on average more polarized.

This modeling finding supports Morgan and Levy (2015)'s claim that generative knowledge and item-specific direct experience trade off gradiently in language processing, such that processing of novel or infrequent items relies upon generative knowledge, with reliance upon item-specific experience increasing with increasing frequency of exposure. Morgan and Levy support this claim with behavioral data, showing that empirical preferences for binomials which are completely novel depend on generative constraints while preferences for frequent expressions depend primarily on frequency of experience with each order. Our modeling results augment this argument by demonstrating that this trade-off is likewise necessary in order to predict the language-wide distribution of preference strengths. In particular, we can conceive of generative knowledge as providing a prior for ordering preferences. Under our final model, the logistic regression component serves an estimate of generative knowledge, which generates preferences clustered unimodally around 0.5. The amount of direct experience one has with an expression then modulates whether it conforms to this prior or whether it deviates. Items with low frequency have a high concentration: they maintain their variability and continue to contribute to the mode around

0.5. Items with high frequency have a low concentration: they are more likely to regularize and contribute to the modes at 0 and 1. Crucially, the inclusion of expression frequency as a predictor of the concentration of the beta distribution is necessary in order to achieve this effect in the model, demonstrating that expressions are indeed relying differentially on generative knowledge versus direct experience depending on their frequency.

This finding fits with previous models of cultural transmission in which, in general, preferences gravitate towards the prior (Griffiths & Kalish, 2005), but with sufficient exposure, exceptions can be learned (e.g. irregular verbs; Lieberman, Michel, Jackson, Tang, & Nowak, 2007). However, this raises a question which is not answered by our or others' models: why don't all expressions converge to their prior preferences eventually? We present two possibilities.

One possibility is that people's probabilistic transmission behavior differs at different frequencies. Convergence to the prior relies upon *probability matching*: people must reproduce variants in approximately the proportion in which they have encountered them. However, this is not the only possible behavior. Another possibility is that people preferentially reproduce the most frequent variant they have encountered, to the exclusion of all other variants, a process known as *regularizing*. If people's tendency to probability match versus regularize is dependent on the frequency of the expression in question (with more regularizing at high frequencies), this could produce the pattern of more polarized expressions at higher frequencies seen in our data. Another possibility is that there is some other unspecified exogenous source of pressure towards regularization, as for instance seems to be the case in child language acquisition (Hudson Kam & Newport, 2009). This pressure might be weak enough that it is overwhelmed by convergence towards the prior at lower frequencies, but can be maintained for items with high enough frequencies to have sufficient exposure to deviate from the prior. Further work is necessary to disentangle these explanations.

In addition to contributing to our understanding of binomial expression processing, we have demonstrated the value of modeling the distribution of idiosyncratic preferences in two ways. First, it has improved our ability to predict preferences for novel items, by better differentiating the rule-following training data from the exceptions. Second, this model turns an observation about language-wide structure (the multimodal distribution of preferences) into a constraint on our theory of the cognitive representation and processing of language (more polarization at higher frequencies).

Acknowledgments

We gratefully acknowledge support from research grants NSF 0953870 and NICHD R01HD065829 and fellowships from the Alfred P. Sloan Foundation and the Center for Advanced Study in the Behavioral Sciences to Roger Levy.

References

Benor, S., & Levy, R. (2006). The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language*, *82*(2), 233–278.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive foundations of interpretation*, 69–94.

The CMU Pronouncing Dictionary . (2014). Carnegie Mellon University.

Cooper, W. E., & Ross, J. R. (1975). World Order. In R. E. Grossman, L. J. San, & T. J. Vance (Eds.), *Papers from the parasession on functionalism* (pp. 63–111). Chicago: Chicago Linguistics Society.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Griffiths, T. L., & Kalish, M. L. (2005, May). A Bayesian view of language evolution by iterated learning. *Proceedings of the 27th annual conference of the cognitive science society*, 827–832.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Levy, R., & Galen, A. (2006). Tregex and Tsurgeon. *5th International Conference on Language Resources and Evaluation (LREC)*.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174.

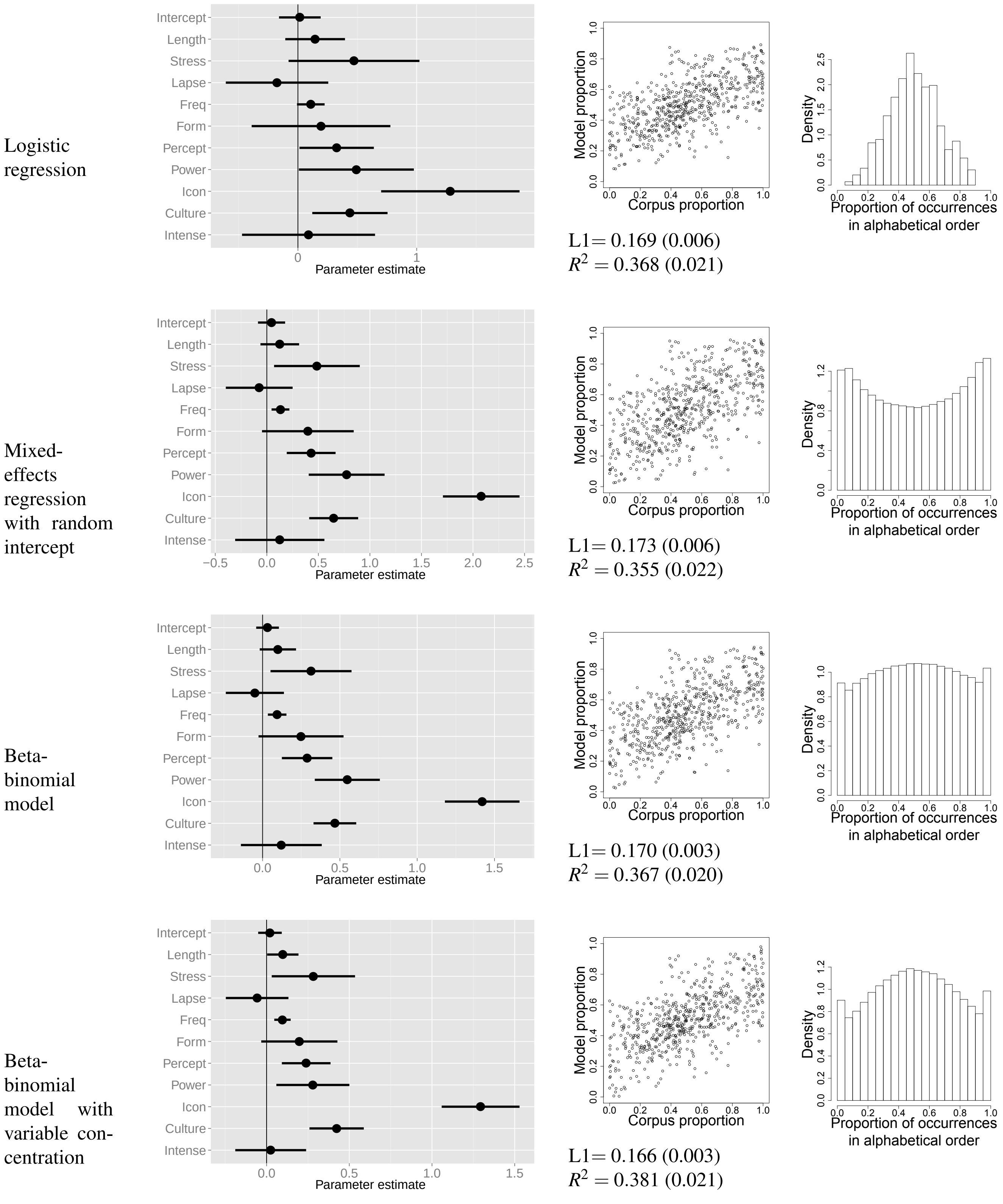
Marcus, M., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3*. Linguistic Data Consortium.

McDonald, J., Bock, K., & Kelly, M. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*, 188–230.

Morgan, E., & Levy, R. (2015). Abstract knowledge versus direct experience in processing of binomial expressions. *Manuscript submitted for publication*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776–784.



(a) (b) (c)

Figure 2: For each of our four models, we display: (a) Parameter estimates for the logistic regression component. Dots show point estimates with bars indicating standard errors. (b) Predictions for each item, as well as mean by-type L1 error and R^2 with by-fold standard errors. (c) Language-wide predicted distribution of preference strengths.