# Brief at the risk of being misunderstood: Consolidating population- and individual-level tendencies

Thomas Brochhagen

thomasbrochhagen@gmail.com

ORCID identifier: 0000-0002-0301-0541

Department of Translation and Language Sciences,
Pompeu Fabra University,
Roc Boronat 138, 08018,
Barcelona, Spain

08.01.2021

**Abstract**

Communicative pressures can give rise to regular patterns of language use. These patterns, in turn, can come to shape a language's structure over time. In a recent study, Kanwal et al. (2017) investigate whether an interaction of such pressures may underlie the cross-linguistic tendency of frequent forms to be shorter. Using a miniature artificial language, they show that speakers follow this tendency if pressured for brevity and accuracy. In this study, we use probabilistic models of varying complexity to shed light on the individual-level factors behind this trend. We find that a hierarchical model that accommodates for subjects' heterogeneous beliefs about object frequencies best explains the data. At the population-level, this model predicts an association of short forms with frequent meanings, in line with past research. At the individual-level, however, it reveals a number of patterns that systematically deviate from this trend. On the one hand, these findings support the hypothesis that individual-level pressures may underlie natural languages' relationship between frequency and brevity. On the other, by characterizing the individual-level dynamics on which this relationship rests, they highlight the importance of consolidating multiple strata of analysis and of understanding where and why they might diverge.

**Keywords:** Zipf's Law of Abbreviation; language universals; rational language use; least effort; ambiguity; efficient communication

## Disclaimer

**This is a post-peer-review, pre-copyedit version of an article published in Computational Brain & Behavior. The final authenticated version is available online at:** https://dx.doi.org/10.1007/s42113-021-00099-x

## Declarations

**Funding:** This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154).

**Conflicts of interest/Competing interests:** None.

**Availability of data and material:** The raw experimental data from Kanwal et al. (2017), which our analysis takes as a starting point, is available at: http://datashare.is.ed.ac.uk/handle/10283/2702.

**Code availability:** All Stan models and analysis code used in this study are available at: https://osf.io/7m9np/?view_only=73729dc33dc7482c84c2f3956ce8eac8.

**Acknowledgments** Thanks to Gemma Boleda, Marco Baroni, Paula Chocron, Matthijs Westera, Robert van Rooij, as well as to two anonymous reviewers and the editor for comments and discussion. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the author's view only, and the EU is not responsible for any use that may be made of the information it contains.

# 1 Introduction

Speakers have to build on limited linguistic resources to convey meaning. Prominently, to form words, natural languages draw from small and closed inventories of discriminatory sounds. This creates an imbalance in the number of possible words of a given length. The shorter the length, the fewer words of that length can there be. In a recent study, Kanwal et al. (2017) investigate how the competition of meanings for such short forms can shape language over time. Their experiments tracked the influence of two pressures on the competition of two meanings –one frequent and the other infrequent– for a single short form. One was a pressure for brevity: Transmission speed depended on sequence length, making the short form preferable to longer alternatives. The second pressure was for communicative accuracy: The short form was ambiguous between the two meanings; longer alternatives were unambiguous. While being faster to send, the short form therefore carried a higher risk of misunderstanding. A comparison across experimental conditions shows that, over time, speakers under both pressures tended to send the short form to signal the more frequent meaning. In other words, the competition of multiple meanings for a finite pool of short and preferred forms may be decided by frequency.

Regularities in the use and acquisition of language at the level of individuals can come to shape language in the longer term (Skyrms 2010, Steels 2011, Tamariz and Kirby 2016). Framed in this bigger picture, Kanwal et al.'s (2017) study provides first experimental support to the idea that a speaker pressure to be accurate but brief might underly Zipf's (1935) *Law of Abbreviation*: the tendency of frequent forms to be short. This tendency is pervasive across natural languages (e.g., Piantadosi et al. 2011, Sigurd et al. 2004, Strauss et al. 2007); as well as other biological signaling (e.g., Ferrer-i-Cancho and Hernández-Fernández 2013, Ferrer-i-Cancho et al. 2013); and can even arise in neural networks' emergent languages when pressured for speaker economy and message discriminability (Chaabouni et al. 2019).

Kanwal et al.'s analysis finds an overall tendency to associate the short form with the frequent meaning. However, this tendency is an abstraction over individual differences. Its identification alone does not shed light on the factors that drive the individual speaker choices on which it is rooted. Understanding these factors and an ensuing consolidation between population trends and individual-level behavior is particularly pressing when dealing with data with much individual variation, as in this case. The present study addresses these issues by means of probabilistic speaker models of varying complexity. We find that subjects' behavior is best explained by a hierarchical model that accommodates for heterogeneous speaker beliefs about their interlocutors' contextual expectations. In line with previous studies, this model predicts a tendency toward the association of the short form with the frequent meaning at the population level. Crucially, at the individual level, it also reveals diverse regularities of language use that go against both Zipf's observed pattern as well as against past individual-level predictions (e.g., Parikh 2000).

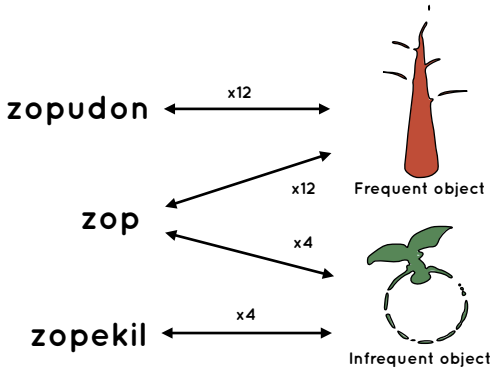


Figure 1: Sketch of a possible language in Kanwal et al. (2017). Edges indicate the number of times an association was witnessed in training.

## 2 Experimental background

Kanwal et al.'s (2017) experiment had two parts. In the first, identical across conditions, subjects were individually trained on a miniature artificial language for 32 trials. This language consists of two "alien" objects and three "alien" words. Two words, *zopudon* and *zopekil*, are long. The third, *zop*, is short. Each training trial showed a licensed alien word-object pair. One object was frequent and appeared in 24 trials. The other only appeared in 8 trials. Each long word was exclusively licensed for one object. For instance, *zopudon* may have only been licensed for the frequent object and *zopekil* only for the infrequent one. By contrast, the short word appeared in half of each object's training trials: 4 times with the infrequent object and 12 times with the frequent one. Figure 1 gives a pictorial summary of the training setup.

After training, subjects were assigned to one of four conditions. The conditions varied depending on whether there was only pressure for communicative success; only for brevity; both; or neither. Pressure for communicative success was enacted by pairing subjects and having them use the language to coordinate on objects for 64 trials. In each of these trials one subject was assigned the role of speaker and the other that of hearer. The speaker was shown one of the two objects and given the choice between sending its long unambiguous name or the short but ambiguous one. The hearer subsequently saw the speaker's chosen word and had to decide between the two objects. They then received feedback on their success, switching roles after each trial. Similarly to training, each speaker had to communicate the frequent object 24 times and the infrequent one 8 times.

In conditions with a pressure for brevity, sending words took time proportional to their character length. To send them, speakers had to hold a transmission button, with characters appearing one at a time. The word was only transmitted once all characters had appeared. Consequently, sending short but ambiguous *zop* took $3/7$ of the time it took to send unambiguous alternatives.

In the condition with both pressures subjects were told that the fastest and most accurate pairs would qualify for a prize. In other words, there was pressure to use the ambiguous message if you wanted to be fast; but there was also a pressure to use the unambiguous messages since they carry less risk of misunderstanding. In conditions lacking an explicit pressure for communicative success, subjects played a single-player naming game. In conditions lacking a pressure for brevity, message transmission was instantaneous.

Since we are interested in the factors that explain the use of an ambiguous but preferred (here: short) word, we focus only on data from the condition with both pressures. Kanwal et al. (2017:50) use a logistic regression with the short name as the binary dependent variable to model this data. Object frequency, trial number, and their interaction were treated as fixed effects, and by-participants intercepts and slopes for object frequency and trial number as random effects. This model predicts a positive interaction between trial number and the short name for the frequent object. That is, in this condition, the short form was associated with the frequent object more as trials progressed. To better understand the linguistic behavior of individuals on which this trend builds, we here fit this data to probabilistic models of varying complexity. In doing so, our goal is not only to estimate the magnitude of this or other linguistic tendencies exhibited by subjects, but also to consolidate heterogeneous individual-level behavior with the population-level tendency past research has mainly focused on. We begin by motivating the individual-level speaker model we build on.

## 3 Analysis

Our goal is to see in how far an individual's use of the short message can be characterized as a function of a pressure for brevity; a pressure for communicative success; latent expectations of object frequencies carried over from training; and dialogal history with a partner –that is, success and failure using the short form to convey a particular meaning. To this end, we fit subjects' speaker data using a probabilistic model of language use (Frank and Goodman 2012, Franke and Jäger 2016, Goodman and Frank 2016). This choice is motivated both by their success in characterizing linguistic phenomena at the semantics-pragmatics interface (see Goodman and Frank 2016 for an overview), as well as by the interpretability of their components in the context of this experiment.

The main idea common to this family of models is that linguistic behavior can be characterized as resulting from reasoning about language use (Grice 1975). Applied to the case at hand, the intuition is that a speaker will use an ambiguous but preferred message only if they reason that their addressee will interpret it as intended. We should clarify, however, that this reasoning procedure does not aim to be a description of the mechanistic process behind subjects' linguistic choices. Even in a relatively artificial setting such as this one, these choices are likely rooted –at least some degree– in leaner and more unconscious and automatic processes. Instead, this family of computational models aims to

give teleological explanations of linguistic behavior (for discussion see Anderson 1990, Jones and Love 2011, Griffiths et al. 2012, Franke and Jäger 2016). That is, they ask whether (boundedly) rational speaker models, couched in an ecological context of interaction, can explain observed linguistic behavior. The temporal nature of this experiment additionally allows us to trace and contrast model predictions of these behaviors over time.

Why would a hearer interpret an ambiguous message one way and not the other? One reason may be that the speaker would have used a different utterance if they meant something else (e.g., Goodman and Stuhlmüller 2013). For example, the utterance *some of my friends are vegan* can be reasoned to imply that *not all are vegan* because, otherwise, the speaker would have used the more informative word *all* instead of *some*. However, in this experiment alternatives to the short form are dispreferred by design. There is no equally good option for either object. Hence, from the sole perspective of speaker-economy it would be best to use the short form for both. Another way to disambiguate is to reason about which referent is more expected in the context of interaction. This idea is echoed across functional analyses of ambiguity. While meaning multiplicity is pervasive in natural language, it does not generally hinder communication. Misunderstandings are avoided because context carries sufficient information for addressees to settle on the intended interpretation (a.o., Piantadosi et al. 2012b, Dautriche 2015, Juba et al. 2011, O'Connor 2015, Brochhagen 2018). Put differently, the short form may be deemed safe if the context is informative enough to clue the hearer in on the intended meaning.

In probabilistic models of language use contextual expectations are standardly represented by a common prior over objects. For instance, if it were true that, after training, speakers expected one object to be three times more likely than the other (Figure 1); and if they additionally believed hearers to have this expectation as well, then they could safely use the short form for the frequent object. Framed in this way, Zipf's *Law of Abbreviation* is what rational language use predicts at the level of individuals as well (e.g., Parikh 2000). Crucially, however, this is true only if all interlocutors have the same contextual expectations; and if these expectations additionally match how often a referent will actually be spoken about. There are good reasons to relax these rather strong assumptions. First, allowing for heterogeneous and private contextual expectations does not imply abandoning the prediction that speakers associate the short form with the frequent meaning. Instead, it acknowledges that there is no evidence, a priori, for assuming that subjects believed their addressees to expect a particular object more than another; nor for these expectations to accord with objects' true frequencies (Brochhagen 2017). By relaxing this assumption, we let the data shed light on this issue. Second, we know from Kanwal et al. that not all subjects associated the short form with the frequent object (see, e.g., Figure 3 in Kanwal et al. 2017). An adequate individual-level model needs to be able to accommodate for such variation both across subjects and across trials. To this end, instead of through a common prior, we model speakers' language use as influenced by their uncertainty about their interlocutor's contextual prior (Brochhagen 2017; 2018). Intuitively, a speaker will send the short but ambiguous message to convey the referent they believe the hearer expects the

most. But, in line with a pressure for communicative success, they will only do so if they are certain enough about these expectations. Otherwise, they will pick the safer but longer form. The behavior of such a speaker $\sigma$, who reasons about a hearer $\rho$, is described by:

$$\rho(r \mid m; pr) \propto L(r, m)\ pr(r), \tag{1}$$

$$\sigma(m \mid s; P) \propto \exp(\lambda((\int P(\theta)\rho(s \mid m; \theta)d\theta) - c(m))), \tag{2}$$

where $\rho(r \mid m; pr)$ is a hearer who interprets a message $m$ as a referent $r$ in proportion to $pr(r)$, the hearer's private contextual expectation over referents, and $L(r, m)$. The latter function codifies the artificial language's lexicon. It returns 1 for licensed referent-message pairs, depicted by arrows in Figure 1, and 0 otherwise. The speaker in (2) reasons about this hearer but does not know their true expectations $pr$. This uncertainty is represented by $P$: the degree to which the speaker believes a prior over referents to match their addressee's expectations. $\theta$ codifies the parameters of $pr$. A soft-max parameter, $\lambda > 0$, models the speaker's rationality (Luce 1959, Sutton and Barto 1998). The higher $\lambda$ is, the more the speaker maximizes communicative success while minimizing the cost function over messages, $c(\cdot)$. In other words, $\lambda$ captures the degree to which a speaker's behavior is sensitive to pressure for communicative success and pressure for brevity. Lower values accordingly correspond to noisier assessments of the hearer's interpretation of a message or of a message's relative cost. For instance, the former could be a consequence of pressure for time: while the experimental design aimed to have speed primarily influence message choice, it may also perturb a speaker's assessment of how their partner will interpret a message.

Beyond $\lambda$, we are interested in estimating subjects' beliefs about their addressees' contextual expectations, captured by $P$. Intuitively, a speaker pressured for brevity and accuracy will name a referent with the short message if they believe their addressee to expect this referent. If they are uncertain, they will use an unambiguous alternative to avoid misunderstandings. Figure 2 illustrates how $P$ can affect what a speaker believes their interlocutor to interpret the short form as.

We let a speaker's beliefs $P$ change based on what is learned about the hearer's latent contextual expectations after each interaction. Since the meaning of the long messages is fixed, only how the short message is interpreted is informative about this matter. Current beliefs $P_{t+1}$ are thus a combination of past beliefs, $P_t$, and the likelihood of a hearer's private expectations, $pr$, leading to the witnessed outcome of interaction $t$ when intending to convey referent $r \in R$ with message $m$:

$$P_{t+1}(pr \mid w(r); m) \propto (\sum_{r' \in w(r)} \rho^0(r' \mid m; pr) P_t(pr)), \tag{3}$$

with $w(r) = \{r\}$ if the interaction was successful and $R - \{r\}$ otherwise. In words, conveying a particular referent with the short message, or failing to do so, gives indirect evidence

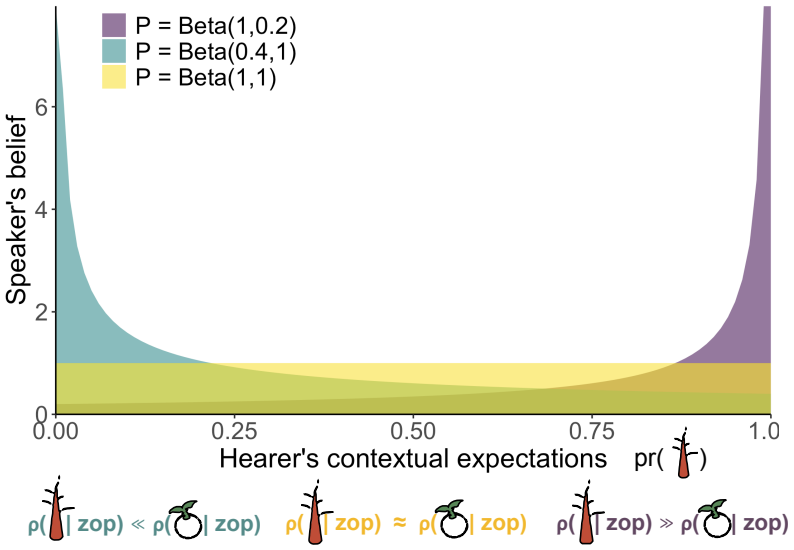


Figure 2: Illustration of speaker beliefs, $P$, about a hearer's contextual expectations $pr$. The more mass there is on higher/lower values of the $x$-axis, the more/less the speaker believes this object to be expected. Consequently, as sketched below the graph, $P$ affects how a speaker believes the short form will be interpreted, $\rho(\cdot \mid \text{zop})$.

about the hearer's private prior over referents. This informs the speaker's future beliefs and, consequently, future uses of this message.

### 3.1 Models

Kanwal et al.'s (2017) condition with both pressures had 40 subjects, with 32 speaker trials per subject. Each of these 1280 data points codifies whether the short name was used; what the intended referent was; and whether the interaction was successful. We use Hamiltonian Monte Carlo, implemented in Stan (Carpenter et al. 2017), to estimate two parameters. The first is $\lambda$, which reflects sensitivity to pressure for communicative success and brevity. The second, $P$, indicates a subject's beliefs about their addressees expectations. We model $P$ as a Beta distribution, parametrized by two shape parameters $\alpha$ and $\beta$. Intuitively, $\alpha$ reflects belief in the frequent referent being expected. Conversely, $\beta$ does so for the infrequent referent (see Figure 2 for illustration).

We fit the data to five models of varying complexity. The first model, NoPool, does not pool the data. It estimates $\lambda_i$, $\alpha_i$ and $\beta_i$ for each subject $i$ separately. On the opposite side of the spectrum, the second model, $\text{FullPool}_{\lambda,\alpha,\beta}$, pools all the data, estimating a single $\lambda$, $\alpha$, and $\beta$. This is tantamount to the assumption that these parameters are shared across subjects. The third, $\text{FullPool}_{\lambda}$ estimates a single pooled $\lambda$; but $\alpha_i$ and $\beta_i$ are estimated separately for each subject $i$, as in NoPool. The last two models are hierarchical.

Differently from NoPool, they allow for individuals' data to inform others' estimates. But, differently from the full pooling models, they also allow for individual variation. They do so by estimating population-level distributions from which the individual-level parameters draw. The first hierarchical model is $\text{HM}_{\lambda}$, with subject $i$'s $\lambda_i$ drawing from a population-level distribution but $\alpha_i$ and $\beta_i$ being individually estimated, as in NoPool. The second and final hierarchical model is $\text{HM}_{\lambda,\alpha,\beta}$. This model additionally assumes individual $\alpha_i$ and $\beta_i$ to be drawn from population-level shape distributions. Appendix A gives an overview of the models' specifications and the priors used.[1]

We expect the pooling models to fare comparably poorly. Data aggregations of this sort are often not good at predicting individual-level data since they are blind to the heterogeneous sources they are tasked to explain (e.g., Estes and Maddox 2005, Franke and Degen 2016). We expect the ability to accommodate for heterogeneous, partner-specific, behavior to be particularly important for this task. The two pooling models are therefore best regarded as baselines to benchmark against. As for NoPool and the two hierarchical models, it is less clear whether the former's relative simplicity is predictively advantageous over the latter two's individual estimates being informed by other subjects' data.

In analogy to Kanwal et al.'s (2017) setup, we fix a nominal preference for the short three-characters long form to $3/7$ of the cost of the two longer ones, both of equal cost: $c_\sigma(\text{zopekil}) = 0.3 = c_\sigma(\text{zopudon})$.

### 3.2 Results

We focus our discussion to models fit with all 1280 data points. Appendices B and C list results for other subsets. The seven subsets in Appendix B were obtained by progressively removing the subjects with the most communicative failures. The two in Appendix C split the data in half, with each split containing data from only one subject per dyad (20 subjects per split). For each fit we ran 4 chains per model for 2000 iterations, with 1000 iterations of warm-up.

**Diagnostics.** All fits were diagnosed to rule out pathologies. These diagnostics all serve the same general purpose: to empirically check whether we arrived at well-behaved Markov Chain Monte Carlo estimators. Plainly put, they indicate whether the results are reliable.

All fits had parameters with a split $\hat{R} < 1.1$ (Gelman and Rubin 1992). This suggests that the chains mixed well, with between- and within-chain estimates agreeing. They also had no saturated trajectory lengths (*max_treedepth* in RStan), meaning that the sampler did not terminate prematurely; no divergent transitions, suggesting no difficulties in exploring the posterior; and an energy Bayesian Fraction of Missing Information over 0.2, indicating no inefficiency in the momentum resampling between trajectories (Betancourt 2017:§6).

[1] Stan models and analysis scripts are available at: https://osf.io/7m9np/?view_only=73729dc33dc7482c84c2f3956ce8eac8.

| | $\text{ELPD}_\Delta$ ($\text{SE}_\Delta$) | ELPD (SE) | EFF (SE) |
|---|---|---|---|
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -471.63 (18.74) | 33.75 (3.77) |
| $\text{HM}_{\lambda}$ | -38.16 (5.74) | -509.79 (18.24) | 25.83 (2.99) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -148.24 (12.17) | -619.88 (22.23) | 3.16 (0.25) |
| $\text{FullPool}_{\alpha,\beta}$ | -154.21 (11.61) | -625.84 (21.24) | 7.83 (0.54) |
| NoPool | -201.02 (34.55) | -672.65 (51.02) | 84.89 (8.58) |

Table 1: Model comparison using approximate leave-one-out cross-validation. $\text{ELPD}_\Delta$ is the difference in expected log point-wise predictive density to the best ranked model. EFF indicates the effective number of parameters.

Another technical challenge faced by Markov Chain Monte Carlo is that samples are usually correlated. It is therefore important to estimate whether the effective sample size is large enough. This was the case for all the fits ($> 0.001$ effective samples per transition).

We use approximate leave-one-out cross-validation to validate our fits as well as for model selection (Vehtari et al. 2017; 2019). All had a shape parameter $k < 0.7$, suggesting reliable leave-one-out estimates.

**Cross-validation and model comparison.** Table 1 shows the models' expected log predictive densities and ranks them.[2] The full hierarchical model, $\text{HM}_{\lambda,\alpha,\beta}$, outperforms both its simpler counterpart $\text{HM}_{\lambda}$ as well as the non-hierarchical models. The same is true of all other data subsets (see Appendices B and C for numeric results). This strongly speaks to the robustness of its relative adequacy and suggests that letting estimates of individuals' parameters be informed and partially constrained by each other improves prediction. And that it does so even in the face of the penalty incurred by the model's relative complexity.

NoPool's predictive acumen is markedly worse than that of both hierarchical models, and quite comparable to that of the pooling models. This result also holds across data subsets and further emphasizes the advantage of the hierarchical models. NoPool's flexibility in estimating each subject's parameters individually results in uncertain estimates that ultimately worsen prediction.

Finally, note that the stark advantage of the hierarchical models holds even if modeling only one subject from each pair (Appendix C). This suggests that population-level trends identified by the models are not just due to the data's latent dyadic structure. Instead, they are regularities common to individuals irrespective to whom they were paired with.

**Estimates.** We focus on the best model, $\text{HM}_{\lambda,\alpha,\beta}$, in what follows. Figure 3 shows posterior estimates for its population-level parameters. At this level, the rationality parameter $\lambda$ suggests a tendency toward a maximization of communicative success while minimizing

[2] The difference in standard error, $\text{SE}_\Delta$, is smaller than individual models' SEs due to correlation.

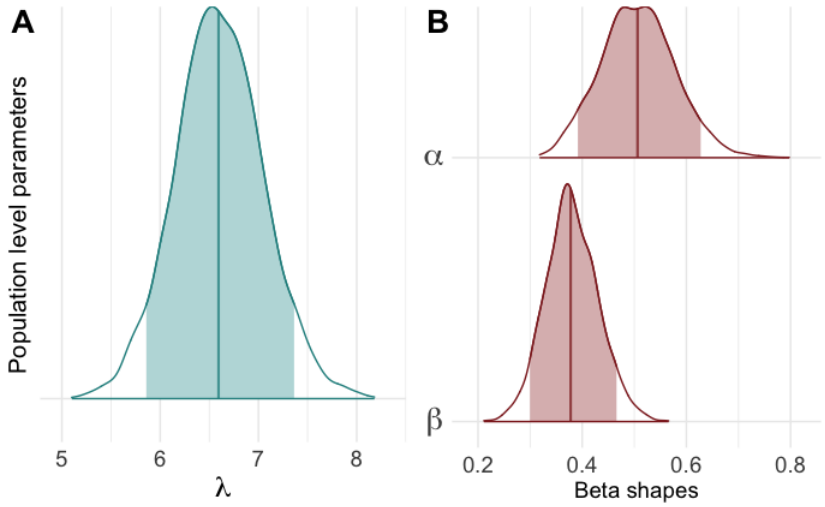


Figure 3: Population parameter estimates. Shadings depict 90% of mass.

message cost. The shapes of the Beta distribution –the initial credence put on the expectation of either of the two objects– favor the frequent object. However, this tendency is not too pronounced. In other words, at the population level speakers' behavior is best explained by a weak initial belief in the frequent object being more expected. These estimates are in line with Kanwal et al.'s (2017) setup and findings.

As shown in Figure 4, however, the individual level reveals striking differences from the population level when it comes to subjects' beliefs and, consequently, their use of the short form. As for individuals' $\lambda$-estimates (Figure 4.A), only a few subjects' linguistic choices are not well explained by sensitivity to pressure for communicative success and brevity. This is consistent with the fact that the trials of five subjects account for more than half of the total communicative failures in this experimental condition. This is also reflected by the close relationship between an individual's expected rationality, $\mathrm{E}[\lambda]$, and their amount of successful trials (Pearson's $r \approx 0.84$). Put differently, the few lower $\lambda$ estimates correspond to subjects that failed to communicate the most. The majority of individual-level $\lambda$ estimates can instead already be intuited by inspecting the population level (Figure 3.A).

Next, we turn to speakers' beliefs about their interlocutors' contextual expectations and how these beliefs changed over time. Initial and final beliefs are shown in Figures 4.B and 4.C, respectively. As hinted at by Figure 3.B, the first thing to note is that most subjects' behavior is initially best characterized by a substantial degree of uncertainty about their interlocutor's expectations. This is a consequence of them experimenting with multiple association patterns for the short form. For example, by first avoiding it but later coming to exclusively associate it with a particular object. As suggested by visual

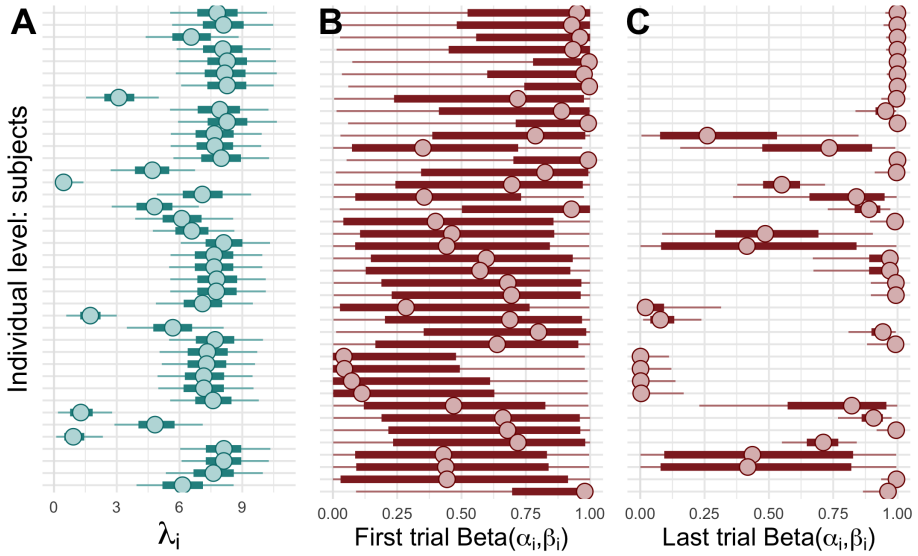


Figure 4: Individual-level parameter estimates. Inner (thicker) intervals depict 50% of mass and outer (thinner) ones depict 90% of mass.

comparison of Figures 4.B and 4.C, this uncertainty decreased over time (Pearson's $r \approx -0.55$ between trial number and width of an individual's $P$ 0.89% highest posterior density interval).

Figure 4.C shows that the short form ended up being used across the entire spectrum of possibilities: from its association with the (in)frequent object to its avoidance. That being said, the belief in the frequent object being more expected (and thus its association with the short form) is still majoritarily represented at the level of individuals. This is so throughout trials. However, it is also true that individuals' beliefs grew away from the population. This departure, measured as the Kullback-Leibler divergence of an individual's beliefs from the initial population-level belief, increases over trials ($r \approx 0.56$). Reversely, the divergence of the beliefs of partners of the same dyad decreased over trials ($r \approx -0.23$) and so did the divergence of an individual's own beliefs from one trial to the next ($r \approx -0.22$). In sum, over time, individuals' certainty about their partner's expectations increase; their expectations grow closer; but they diverge more from the initial population-level beliefs. These findings are intimately related. Subjects began interacting with substantial uncertainty about their interlocutors beliefs (Figure 4.B). This is also reflected by the population-level estimate in Figure 3.B. As they interact, however, interlocutors grow closer and more certain about each other's expectations. This consequently implies a departure from the initial population belief. Notably, a number of subjects did not follow Zipf's and Kanwal et al.'s (2017) predicted pattern. Instead, they came to associate the short form with the infrequent object or to not use it at all (subjects in Figure 4.C with substantial mass on lower or

central $x$-values, respectively; see *Predictions* below for further illustration).

As with $\lambda$, some changes in an individual's beliefs are also related to their communicative success:[3] the divergence of an individual's beliefs to their partner's is negatively correlated to their success rate ($r \approx -0.21$) and so it is also to the divergence of an individual's own beliefs from one trial to the next ($r \approx -0.21$). However, neither an individual's divergence to the population belief nor the width of the 0.89% highest posterior density interval of their beliefs are related to their rate of success ($r \approx 0.004$ and $r \approx 0.02$, respectively). This is explained by the fact that a plethora of different beliefs can lead to communicative success. As suggested by Figure 4, these need not agree with the population-level belief from which they initially draw from, nor need they be particularly narrow, as is the case for subjects that avoided the use of the short form due to uncertainty about their interlocutors expectations. Lastly, by contrast to $\lambda$, beliefs can change after both success and failure. This makes their relationship to communicative success more indirect.

**Predictions.** Figure 5 shows a selection of posterior predictive checks. The upper row shows a subject that always succeeded in using the short form for the frequent object (left) and one that did so for the infrequent object (right). They correspond to two of the opposite extremes in Figures 4.B and 4.C. That is, their behavior is well explained by opposite beliefs about their interlocutor's latent contextual expectations. These beliefs are reinforced by trials in which the short form is successfully interpreted as the intended object. The lower row of Figure 5 shows a subject that never used the short form (left) and a subject that experienced 10 communicative failures (right). While the former's behavior is estimated to arise from deeming the short form too risky to use; the latter's, somewhat erratic, behavior remains largely unaccounted for by the model.

In sum, the best model identifies and can account for behavior that varies along three communicative patterns: a tendency to associate the short form with the frequent meaning; with the infrequent meaning; or its avoidance. The model's predictive success decreases with the amount of communicative failures experienced. This is expected given that we assume subjects to be sensitive to the experiment's communicative pressures. Particularly, that they primarily cared about communicative success.[4]

As for overall predictive accuracy, $\text{HM}_{\lambda,\alpha,\beta}$ has a root-mean-square error of 0.32, measured as the root of the mean difference between actual values, $y$, and the ones predicted, $\hat{y}$. In light of the categorical and heterogeneous nature of this data, this is a good performance. For comparison, predicting the short form to always go with the frequent object scores a

[3]Note that we quantified the relationship between an individual's total success and their $\lambda$-estimate. Since belief-related measures change, we instead compare them to individuals' rate of success.

[4]This is a standard assumption in both the game-theoretic (Blume et al. 1993, Benz and van Rooij 2007) and the probabilistic pragmatics literature (Qing and Franke 2015). While it is possible that some subjects cared more about communicating as fast as possible, even in the face of repeated failure, or that they did not care about communicative success at all, we keep this simplifying assumption for lack of a principled alternative.

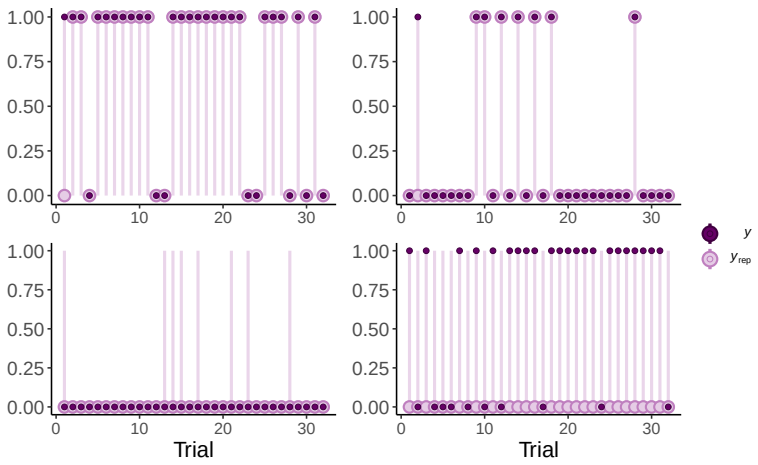


Figure 5: Posterior predictive for four speakers. The $y$-axis shows use (1) or avoidance (0) of the short form, indicated by $y$, against model predictions $\hat{y}$, colored with 85% mass.

root-mean-square error of 0.63; always associating it with the infrequent object 0.78; and always avoiding it 0.67. Since subjects explored different patterns of use, accounting for their choices at every trial is challenging. The root-mean-square error expectedly decreases for models fit on data that excludes subjects that had the most unsuccessful trials. For instance, it lowers to 0.28 when excluding the 4 worst-faring subjects; and to 0.24 when excluding the 8 worst-faring ones.

## 4 Discussion

In line with Kanwal et al.'s (2017) findings, the preceding analysis shows that subjects' behavior, as a whole, is best characterized by a tendency to associate the short form with the frequent meaning. This tendency may be rooted in an initial belief that the frequent object is more expected (Figure 3.B). This belief is formed and carried over from the experiment's training phase. Consequently, the short but ambiguous form is majoritarily taken to be interpreted as the frequent object. This result underscores the influence that skewed meaning frequencies can have on an evolving linguistic system (e.g., Perfors and Navarro 2014, Brochhagen 2018).

When it comes to the individual-level behavior behind this trend, systematic divergences from the population level crystallize. In particular, both an avoidance of the short but ambiguous form as well as its association with the infrequent meaning are well-attested. Our analysis suggests that different patterns of use can be traced back to differences in the beliefs that subjects form about each other's latent contextual expectations. These findings add to the growing number of studies that show how population-level tendencies need not

be reflected by the individuals that constitute them (e.g., Franke and Degen 2016, Brochhagen 2018). More narrowly, this result vindicates behaviors that previous analyses ruled out in virtue of their suboptimality (e.g., Parikh 2000, Bergen et al. 2012) or missed by focusing only on the population (Kanwal et al. 2017). It shows how weakening the assumption of a common contextual prior allows the data itself to speak to whether individuals associate frequent meanings to preferred forms. As it turns out, a non-negligible minority of speakers did not follow this pattern. More broadly, over time and across association patterns, dyads' beliefs grew closer; subjects became more certain about their partner's expectations; their individual beliefs diverged from the initial population-level belief; and their use of the short form became more deterministic. Taken together, these changes can be regarded as signatures of the mutually reinforcing effect of repeated linguistic interactions.

These results support the hypothesis that frequency has a bearing on the likelihood of a form being associated with a meaning. In particular, that preferred forms will tend to be associated with frequent meanings. However, this trend is not inevitable. Instead, micro-dynamics at the level of dyads can lead to other association patterns. That is, the initial influence of frequency can be overcome and partners sometimes settle on regularities that, while not most optimal, work well enough: they are a good match for the expected and, over time, entrenched patterns of use established with one another.

How and when, then, do association patterns established by dyads find their way into the population? This is a very important question that, while outside the scope of the data we are analyzing, should still be touched upon. Our analysis suggests that not all possible association patterns are equally likely: the optimal one has a larger basin of attraction. How large this basin is can be hypothesized to depend on the difference between the frequencies competing for a single form. Investigating the relationship between this difference and the likelihood that a form will come to be associated to one of many candidate meanings is a promising venue for future research, calling for both diachronic corpus analysis and further laboratory experimentation. Besides communicative pressure for success and brevity, other factors such as learnability and memory may also tilt the field in favor of the *Law of Brevity*: A meaning that is more frequent has more chances to be associated with a form; to be remembered; and to be faithfully passed on. In sum, while suboptimal associations might occasionally get entrenched simply because using a longer expression is disprefered to abandoning a previously successful association pattern, on average, differences in meaning frequencies will tend to favor the optimal pattern.

However, these considerations still abstract away from the dynamics at play when it comes to relating the individual-level back to the population. A concrete and recent proposal to this end is found in Hawkins et al. 2020. This work shows that a hierarchical model that feeds dyadic conventions back into community-level expectations over them fares well in characterizing the way in which local linguistic knowledge can reach beyond partner-specific boundaries. This idea is fully compatible with our hierarchical models. However, new data is needed to evaluate it in this scenario. In particular, data on speakers changing their partners is needed in order to study the degree to which beliefs and

association patterns are carried over to novel partners.

Lastly, some words of caution on the interpretation of these results are due. Just as this data alone does not speak to how an association pattern may find its way into the population, it is also not possible to speak to the nature of the form-meaning associations that subjects established. On the one hand, one might argue that interlocutors optimized their lexica such that the short form came to be literally associated with only one object (Kanwal et al. 2017). That is, this view argues that these associations are long lasting and semantic in nature. On the other, it is equally plausible that these associations were opportunistic outcomes of a partner-specific adaptation enabled by the context of interaction (e.g., Clark and Wilkes-Gibbs 1986, Metzing and Brennan 2003, Brennan and Clark 1996); or that they draw from general pragmatic principles. For instance, where Horn's (1984) division of labor predicts (un)marked forms to receive (un)marked interpretations, it would here predict ambiguous forms to receive the most predictable meaning (a.o., Piantadosi et al. 2012a, Dautriche 2015, Juba et al. 2011, Brochhagen 2017). That is, one might instead argue that subjects' associations reflect pragmatic refinings, suggesting that the underspecified meaning of the short form remained intact. Our probabilistic speaker model instantiates this second, pragmatic, view. However, we do not intend its relative success to be taken as support for this view. Since the experimental setup does not speak to this matter, the relevant modelling choices should be viewed as necessary assumptions rather than as principled choices. We hope these open issues invite future research on the lexicalization of frequency-driven phenomena at the semantics-pragmatics interface.

## 5 Conclusion

Recent years have seen a steady stream of studies that seek to answer how communicative pressures and the context in which language is used interact and come to shape it. The goal of this study was to add to this broad enterprise by investigating individual-level dynamics that underly the tendency of short forms to be associated with frequent meanings. To this end, we employed a probabilistic speaker model that factors in uncertainty about interlocutors' contextual expectations. Such latent (un)certainty was shown to succeed in characterizing the use of a short but ambiguous form when unambiguous alternatives are an option. Furthermore, we showed that Kanwal et al.'s (2017) data is best explained by a hierarchical model that estimates population-level parameters from which the individual level draws. This study thereby consolidates two levels that had previously only been studied disjointly, individuals and the population they make up, and highlights where they diverge.

Our findings suggest that exposure to skewed meaning frequencies can indeed give rise to a tendency for short forms to be associated with frequent meanings. This might be due to their relative predictability (Piantadosi et al. 2012a, Dautriche 2015, Juba et al. 2011), which is reinforced the more interlocutors successfully use them (Brochhagen 2017).

However, at the level of individuals, we also identified a number of behavioral patterns that go against this regularity. It is important to not gloss over these suboptimal alternatives. First, because an adequate explanation of how conventions reached by dyads find their way into a population's language needs to account for them. Second, because they should caution against ruling out behaviors deemed suboptimal. Echoing Franke and Degen (2016): focusing only on population-level trends may hide patterns at the individual level that ultimately contribute to them.

# Appendices

## A Models

Specifications of models introduced in Section 3.1 together with priors, truncated from below at 0, employed in the analyses in Section 3.2. All models are written in Stan (Carpenter et al. 2017). They are available with accompanying analysis scripts at: https://osf.io/7m9np/?view_only=73729dc33dc7482c84c2f3956ce8eac8.

$$
\begin{aligned}
\textsc{NoPool} \qquad & \lambda_i \sim \text{Normal}(20, 1.5) \\
& \alpha_i \sim \text{Normal}(1, 0.2) \\
& \beta_i \sim \text{Normal}(1, 0.2)
\end{aligned}
$$

$$
\begin{aligned}
\textsc{FullPool}_{\lambda,\alpha,\beta} \qquad & \lambda \sim \text{Normal}(20, 2) \\
& \alpha \sim \text{Normal}(1, 0.2) \\
& \beta \sim \text{Normal}(1, 0.2)
\end{aligned}
$$

$$
\begin{aligned}
\textsc{FullPool}_{\lambda} \qquad & \lambda \sim \text{Normal}(20, 2) \\
& \alpha_i \sim \text{Normal}(1, 0.2) \\
& \beta_i \sim \text{Normal}(1, 0.2)
\end{aligned}
$$

$$
\begin{aligned}
\text{HM}_{\lambda} \qquad & \mu_\lambda \sim \text{Normal}(15, 5) \\
& \lambda_i \sim \text{Normal}(\mu_\lambda, 2) \\
& \alpha_i \sim \text{Normal}(1, 0.2) \\
& \beta_i \sim \text{Normal}(1, 0.2)
\end{aligned}
$$

$$
\begin{aligned}
\text{HM}_{\lambda,\beta,\alpha} \qquad & \mu_\lambda \sim \text{Normal}(15, 5) \\
& \mu_\alpha \sim \text{Normal}(1, 0.2)
\end{aligned}
$$

$$\mu_\beta \sim \text{Normal}(1, 0.2)$$
$$\lambda_i \sim \text{Normal}(\mu_\lambda, 1.5)$$
$$\alpha_i \sim \text{Normal}(\mu_\alpha, 0.2)$$
$$\beta_i \sim \text{Normal}(\mu_\beta, 0.2)$$

# B Leave-one-out cross-validations subsetted by success

Results of approximate leave-one-out cross-validations for varying data subsets. Subsets were created by progressively including only subjects that had at most $n$ unsuccessful trials, with $n \in [1, 9]$. We only report on subsets made up of different amounts of subjects. For example, there are two subjects that had more than 9 unsuccessful speaker trials. No additional subjects had more than 8 unsuccesful trials, meaning that $n = 9$ and $n = 8$ yield the same subset.

| | $\text{ELPD}_\Delta$ ($\text{SE}_\Delta$) | ELPD (SE) | EFF (SE) |
|---|---|---|---|
| **< 10 unsuccessful trials (38 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -416.78 (20.32) | 36.90 (4.46) |
| $\text{HM}_{\lambda}$ | -41.64 (6.67) | -458.42 (18.98) | 27.19 (3.33) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -113.08 (11.33) | -529.86 (24.05) | 3.89 (0.41) |
| $\text{FullPool}_{\lambda}$ | -118.37 (10.61) | -535.15 (22.67) | 11.42 (0.95) |
| NoPool | -154.57 (30.20) | -571.35 (48.47) | 75.79 (8.28) |
| **< 8 unsuccessful trials (36 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -363.96 (21.27) | 38.56 (5.02) |
| $\text{HM}_{\lambda}$ | -43.81 (7.58) | -407.76 (19.24) | 27.46 (3.45) |
| $\text{FullPool}_{\lambda}$ | -95.78 (10.65) | -459.74 (24.14) | 15.22 (1.45) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -97.36 (11.42) | -461.32 (25.43) | 4.75 (0.61) |
| NoPool | -113.47 (24.74) | -477.42 (44.00) | 67.47 (7.91) |
| **< 7 unsuccessful trials (35 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -333.69 (23.63) | 42.91 (5.93) |
| $\text{HM}_{\lambda}$ | -46.26 (9.05) | -379.95 (20.08 ) | 28.89 (3.72) |
| $\text{FullPool}_{\lambda}$ | -70.42 (8.08) | -404.11 (24.89) | 20.28 (2.14) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -78.47 (9.51) | -412.16 (26.13) | 5.05 (0.73) |
| NoPool | -83.85 (19.30) | -417.54 (41.32) | 62.14 (7.66) |
| **< 5 unsuccessful trials (34 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -326.32 (23.26 ) | 41.80 (5.85) |
| $\text{HM}_{\lambda}$ | -48.47 (9.13) | -374.79 (19.65) | 28.30 (3.70) |

*(continued)*

| | $\text{ELPD}_\Delta$ ($\text{SE}_\Delta$) | ELPD (SE) | EFF (SE) |
|---|---|---|---|
| $\text{FullPool}_\lambda$ | -72.53 (8.06) | -398.85 (24.50) | 19.66 (2.08) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -78.17 (9.37) | -404.49 (25.81) | 4.82 (0.69) |
| NoPool | -84.07 (19.17) | -410.39 (40.80) | 60.13 (7.53) |
| **< 4 unsuccessful trials (32 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -267.51 (24.80) | 43.97 (6.90) |
| $\text{HM}_\lambda$ | -51.27 (9.48) | -318.77 (20.88) | 29.88 (4.27) |
| NoPool | -55.91 (12.08) | -323.41 (34.46) | 52.65 (7.22) |
| $\text{FullPool}_\lambda$ | -57.00 (7.75) | -324.50 (24.34) | 26.35 (3.18) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -73.06 (10.32) | -340.56 (25.14) | 5.34 (0.85) |
| **< 3 unsuccessful trials (29 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -233.70 (23.71) | 38.17 (6.11) |
| $\text{HM}_\lambda$ | -42.81 (8.73) | -276.51 (20.27) | 28.31 (4.36) |
| NoPool | -46.36 (10.88) | -280.06 (32.37) | 45.57 (6.52) |
| $\text{FullPool}_\lambda$ | -49.08 (7.13) | -282.78 (24.09) | 25.40 (3.31) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -66.93 (10.26) | -300.63 (24.55) | 5.54 (1.00) |
| **< 2 unsuccessful trials (25 subjects)** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -214.03 (21.95) | 33.49 (5.64) |
| $\text{HM}_\lambda$ | -35.30 (7.82) | -249.33 (18.72) | 23.88 (3.94) |
| $\text{FullPool}_\lambda$ | -41.75 (6.42) | -255.78 (22.58) | 21.84 (2.99) |
| NoPool | -43.80 (11.47) | -257.84 (31.71) | 41.27 (6.24) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -60.20 (9.17) | -274.23 (22.51) | 5.06 (0.89) |

Table 2: Model comparisons using approximate leave-one-out cross-validation. $\text{ELPD}_\Delta$ is the difference in expected log pointwise predictive densities to the best ranked model. EFF indicates the effective number of parameters.

# C Leave-one-out cross-validations of dyad splits

Results of approximate leave-one-out cross-validations for two splits of the data. Each split includes only one member of each dyad. That is, if Alice and Bob formed a dyad, Alice is assigned to *Split A* and Bob to *Split B*. Consequently, each split collects data from 20 individuals (640 data points).

| | $\text{ELPD}_\Delta$ ($\text{SE}_\Delta$) | ELPD (SE) | EFF (SE) |
|---|---|---|---|
| **Split A** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -212.34 (15.54) | 19.98 (3.71) |
| $\text{HM}_{\lambda}$ | -32.20 (4.99) | -244.54 (13.89) | 11.98 (2.05) |
| NoPool | -70.83 (19.47) | -283.17 (34.03) | 39.29 (6.14) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -70.91 (9.36) | -283.25 (16.18) | 3.34 (0.43) |
| $\text{FullPool}_{\lambda}$ | -74.20 (9.13) | -286.54 (14.98) | 5.01 (0.53) |
| **Split B** | | | |
| $\text{HM}_{\lambda,\alpha,\beta}$ | 0.00 (0.00) | -262.06 (11.66) | 14.55 (1.96) |
| $\text{HM}_{\lambda}$ | -17.25 (2.62) | -279.32 (11.26) | 9.89 (1.29) |
| $\text{FullPool}_{\lambda}$ | -74.80 (6.81) | -336.87 (13.59) | 3.71 (0.32) |
| $\text{FullPool}_{\lambda,\alpha,\beta}$ | -79.24 (6.92) | -341.30 (13.55) | 2.31 (0.25) |
| NoPool | -127.44 (28.16) | -389.51 (37.88) | 45.18 (5.80) |

Table 3: Model comparisons using approximate leave-one-out cross-validation. $\text{ELPD}_\Delta$ is the difference in expected log pointwise predictive densities to the best ranked model. EFF indicates the effective number of parameters.

## References


J. R. Anderson. *The Adaptive Character of Thought*. Psychology Press, 1990.

A. Benz and R. van Rooij. Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi*, 26(1):63–78, 2007. doi: 10.1007/s11245-006-9007-3.

L. Bergen, N. D. Goodman, and R. Levy. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of 34th Annual Meeting of the Cognitive Science Society*, 2012.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

A. Blume, Y.-G. Kim, and J. Sobel. Evolutionary stability in games of communication. *Games and Economic Behavior*, 5(4):547–575, 1993. doi: 10.1006/game.1993.1031.

S. E. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996. doi: 10.1037/0278-7393.22.6.1482.

T. Brochhagen. Signaling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science*, 2017. doi: 10.1093/bjps/axx058.

T. Brochhagen. *Signaling under Uncertainty*. PhD thesis, University of Amsterdam, 2018.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017. doi: 10.18637/jss.v076.i01.

R. Chaabouni, E. Kharitonov, E. Dupoux, and M. Baroni. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems 32*, pages 6293–6303. Curran Associates, Inc., 2019.

H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1): 1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.

I. Dautriche. *Weaving an Ambiguous Lexicon*. PhD thesis, École Normale Supérieure, 2015.

W. K. Estes and W. T. Maddox. Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, 12(3):403–408, 2005. doi: 10.3758/bf03193784.

R. Ferrer-i-Cancho and A. Hernández-Fernández. The failure of the law of brevity in two new world primates. statistical caveats. *Glottotheory*, 4(1), 2013. doi: 10.1524/glot.2013.0004.

R. Ferrer-i-Cancho, A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M. J. Hsu, and S. Semple. Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578, 2013. doi: 10.1111/cogs.12061.

M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633.

M. Franke and J. Degen. Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PLoS ONE*, 11(5), 2016. doi: 10.1371/journal.pone.0154854.

M. Franke and G. Jäger. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1), 2016. doi: 10.1515/zfs-2016-0002.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136.

N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016. doi: 10.1016/j.tics.2016.08.005.

N. D. Goodman and A. Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5:173–184, 2013. doi: 10.1111/tops.12007.

P. Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

T. L. Griffiths, N. Chater, D. Norris, and A. Pouget. How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, 138(3):415–422, 2012. doi: 10.1037/a0026884.

R. D. Hawkins, N. D. Goodman, A. E. Goldberg, and T. L. Griffiths. Generalizing meanings from partners to populations: Hierarchical inference supports convention formation on networks. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

L. R. Horn. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin, editor, *Meaning, Form and Use in Context*, pages 11 – 42. Georgetown University Press, 1984.

M. Jones and B. C. Love. Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169–188, 2011. doi: 10.1017/s0140525x10003134.

B. Juba, A. T. Kalai, S. Khanna, and M. Sudan. Compression without a common prior: An information-theoretic justification for ambiguity in language. In *Proceedings of the 2nd Symposium on innovations in computer science*, 2011.

J. Kanwal, K. Smith, J. Culbertson, and S. Kirby. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52, 2017. doi: 10.1016/j.cognition.2017.05.001.

D. R. Luce. *Individual choice behavior: a theoretical analysis*. Wiley, 1959.

C. Metzing and S. E. Brennan. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2): 201–213, 2003. doi: 10.1016/s0749-596x(03)00028-7.

C. O'Connor. Ambiguity is kinda good sometimes. *Philosophy of Science*, 82(1):pp. 110–121, 2015. doi: 10.1086/679180.

P. Parikh. Communication, meaning, and interpretation. *Linguistics and Philosophy*, 23 (2):185–212, 2000.

A. Perfors and D. J. Navarro. Language evolution can be shaped by the structure of the world. *Cognitive Science*, 38(4):775–793, 2014. doi: 10.1111/cogs.12102.

S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011. doi: 10.1073/pnas.1012551108.

S. T. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012a. doi: 10.1016/j.cognition.2011.10.004.

S. T. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012b. doi: 10.1016/j.cognition.2011.10.004.

C. Qing and M. Franke. Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics*, pages 201–220. Springer International Publishing, 2015. doi: 10.1007/978-3-319-17064-0_9.

B. Sigurd, M. Eeg-Olofsson, and J. van Weijer. Word length, sentence length and frequency - zipf revisited. *Studia Linguistica*, 58(1):37–52, 2004. doi: 10.1111/j.0039-3193.2004.00109.x.

B. Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, 2010.

L. Steels. Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356, 2011.

U. Strauss, P. Grzybek, and G. Altmann. Word length and word frequency. In *Contributions to the Science of Text and Language*, pages 277–294. Springer Netherlands, 2007. doi: 10.1007/978-1-4020-4068-9_13.

R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1998.

M. Tamariz and S. Kirby. The cultural evolution of language. *Current Opinion in Psychology*, 8:37–43, 2016.

A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432, 2017.

A. Vehtari, J. Gabry, M. Magnusson, Y. Yao, and A. Gelman. loo: Efficient leave-one-out cross-validation and WAIC for bayesian models, 2019. URL https://mc-stan.org/loo. R package version 2.2.0.

G. Zipf. *The Psycho-Biology of Language*. Houghton Mifflin, 1935.