Exploring a pragmatic account of Zipf's Law of Abbreviation using Bayesian data analysis

Benjamin N. Peloquin
bpeloqui@stanford.edu
Department of Psychology
Stanford University

Joseph Cornelius
jocorn@stanford.edu
Department of Psychology
Stanford University

Noah D. Goodman
noah@university.edu
Department of Computer Science
Stanford University

Michael C. Frank
mike@university.edu
Department of Psychology
Stanford University

Abstract

Zipf (1935) argued that the relation between a word's frequency and length was a universal property of human language. This relation, what Zipf called the "Law of Abbreviation", emerged from the competing pressures to minimize speaker- and listener-effort during communication. Subsequent projects have primarily adopted methodological tools in line with Zipf's original work – relying on analyses of large-scale, observational data. Such analyses, however, can only provide indirect evidence for the causal role speaker- and listener-pressures have on the "Law of Abbreviation." In an effort to isolate these competing dynamics experimentally, Kanwal et al. (2017) conducted a study using an artificial language learning paradigm. Only when both a speaker- and listener-pressures were present in a communication game, did the "Law of Abbreviation" emerge. Crucially, the authors highlight that their data is consistent with two possible analyses – a Zipfian account and a pragmatic, language use account – but that it is impossible to distinguish between the two given nature of the data collected. In this work, we consider the pragmatic, language-use account by conducting a Bayesian data analysis, modeling subjects in Kanwal et al. (2017) using the computational-pragmatics framework. Fitting subject-level parameters we find good fit between model posterior-predictive values and experimental production data. We consider these results in light of arguments put forth in Kanwal et al. (2017) with respect to the relation between pragmatic language use and large-scale language change.

Keywords: Language change; Pragmatics; Bayesian data analysis; Bayesian cognitive modeling

Introduction

Zipf (1935) presented a view of human behavior that focused on *effort minimization* as a guiding principle. Relying on empirical evidence from language corpora, he argued that the unique dynamics of speaker- and listener-effort minimization give rise to the distributional forms found in natural language. Among these, Zipf examined the relation between a word's frequency in a corpus and a variety of properties including its rank-frequency, its denotation size, and its length. In terms of this last property, what he called the "Law of Abbreviation," (LOA) Zipf claimed that the more frequent words are, the shorter they tend to be. Evidence for this relationship can be found across languages (CITATION), may also be present in animal communication systems (Ferrer-i-Cancho et, 2013), and has also been observed in artificial (programming) languages (CITATION from Kanwal).

Zipf argued that this relationship between frequency and length was an emergent property of competing speaker-listener pressures. Under this framing, communication is viewed as a cooperative, joint activity between interlocutors. However, this joint activity contains a fundamental asymmetry – what is effortful for a speaker (production) is different from what is effortful for a listener (understanding). An optimal speaker language, Zipf argued, would consist of a single, low-cost word which was fully ambiguous; referring to all possible objects in the world. Adopting such a language, a listener would need to disambiguate every utterance by the speaker. By contrast, if a language were optimised solely in terms of listener effort, the language should bijectively map all forms to meanings so there was no need for a listener to disambiguate.

Zipf argued that these competing forces, what he called "Speaker's economy" and "Auditor's economy", give rise to the particular distributional forms found in all languages, including the relationship between word frequency and length. While this framework is compelling and can largely be seen as foundational to a family of theories examining efficiency in language-structure (CITATIONS) and -use (CITATIONS), the issue is far from settled. Subsequent work has argued that properties such as the relation between word frequency and length can theoretically be explained by a process of random typing (Ferrer-i-Cancho & Mascoso del Pardo, 2012). Kanwal et al. (2017) highlight the fact that Zipf's original work, as well as more recent studies analysing large-scale, observational corpora, "do not explicitly target the hypothesized role of communicative pressures…" In other words, it is difficult to make causal claims about the role that competing speaker- and listener-pressures play purely from large-scale, observational data analysis.

To gain traction on the causal role of speaker-listener pressures, Kanwal et al. (2017) implemented an artificial language learning paradigm to isolate the pressures directly. In line with Zipf's original analysis, they hypothesized that only when *both* speaker- and listener- pressures are present in a communication setting would the LOA emerge. The authors adopted a miniature artificial language learning set-up in which participants learned a language consisting of three words, which could be used to refer to two items. Crucially, two of the words were longer (seven versus three characters) than a third word. The two longer words could only be used to refer to one of the two objects. The third word, an abbreviation of the longer forms, could be used to refer to either of the two objects (it was ambiguous).

Operationalizing production cost as a function of word length, they hypothesized that when given a choice between alternatives, subjects should prefer to use the shorter, less costly term to refer to a more frequently occurring object. Importantly, however, this choice should only appear when

there were both speaker- and listener-pressures present. Results indicated that speakers did tend to use the less costly form for the more frequent item when both pressures were present, but, importantly, not in conditions where only one or neither of the pressures were present.

While Kanwal et al. (2017) represents an important experimental step in confirming Zipfian notions about lexicon-level efficiency, the authors highlight an important and subtle dimension to these results. Zipf's LOA is a lexicon-level theory. That is, it attempts to explain why particular word forms are mapped to particular meanings. However, their experimental results are, in theory, consistent with a pragmatic language-use account as well. That is, the underlying lexicon learned by subjects may have remained unchanged during the experiment – a finding which, on its face, is at odds with the claim that this experiment demonstrates lexicon-level change. Rather, subjects may have *used* the lexicon pragmatically, preferring to map less costly forms to more frequent objects, while leaving the underlying lexicon unchanged. The authors' highlight that because they only recorded participants' actual language production they cannot distinguish between the lexicon-change account and pragmatic, language-use account.

The current project addresses the question of whether patterns of production behavior observed in Kanwal et al. (2017) can be explained by a pragmatic, language-use account. To do so, we conduct a Bayesian data analysis, modeling subjects as rational pragmatic agents using the Rational Speech-Act framework (RSA) (Frank & Goodman, 2017; Goodman & Frank, 2016). To preview our results we find good fit from model posterior-predictive values to Kanwal et al. (2017) participant data ($r^2 = 0.972$), over and above what would would be expected from a baseline model ($r^2 = 0.466$). We believe this provides compelling evidence in support of the alternative hypothesis outlined by Kanwal et al. (2017) – that subjects may have been behaving pragmatically, rather than demonstrating real change at the level of language-structure.

The paper proceeds as follows. We describe the experimental set-up of Kanwal et al. (2017), their hypotheses and the alternative interpretation of their results. We then introduce our general approach to the Bayesian data analysis as well as the Rational Speech-act framework we will use to model subject's pragmatic language use. We evaluate our results summarising the findings and highlighting dimensions of their analysis our current implementation does not yet capture. Following this work we return the question posed by Kanwal et al. (2017) of how pragmatic language use might bootstrap language change and offer an argument for how this type of process could in theory unfold.

Kanwal et al. (2017) and experimental evidence for Zipf's Law of Abbreviation

Artificial language learning paradigms have proven to be a productive methodology for studying language change in the laboratory (or online via web experiments) (CITATIONS). In the typical set-up, participants are trained on a simple language – a series of form-meaning mappings, often using nonce words with novel objects. During test, participant's language production is recorded, typically over a series of communication events or across multiple generations of participants accomplishing some task (CITATIONS). Subsequent analyses focus on how properties of the produced language change as it is used over time.

Kanwal et al. (2017) adopt this basic framework – participants learned a miniature artificial language and were subsequently asked to communicate with an interlocutor in a reference game setting. Following ideas first indicated by Zipf (1935), the authors hypothesized the presence or absence of pressures to communicate quickly (speaker pressure) and accurately (listener pressure) should impact the resulting languages. Participants were assigned to one of four possible conditions which varied in the pressures that were present. Both speaker- and listener-pressures were present in the *combined* condition, only the listener pressure present in the *accuracy* condition, and only the speaker pressure present in the *time* condition. They also tested a fourth *neither* condition in which neither speaker nor listener pressures were present. For the purpose of our current project, we do not report our results on the control, *neither* condition, as behavior should be arbitrary.

For a comprehensive description of the experimental methodology we refer the reader to Kanwal et al. (2017), focusing on essential details here. During training participants learned labels for two novel objects. Importantly, one of the objects occurred three times as frequently than the other (appearing 24 vs 8 times). During training participants observed the objects at their relative frequencies as well as labels. Labels appeared either as a long-form (e.g. "zopudon" or "zopekil") or short form ("zop"). Importantly, in this set-up "zop" was ambiguous between the two object referents. During test, participants were asked to participate in a communication game in which a "director" was shown an object and asked to transmit its label to the "matcher." The "director" was always given a choice to use the long-form label or the ambiguous short-form. Crucially, to transmit the label to the matcher the director had to click on it, holding the mouse while each character appeared sequentially. In this setting, transmitting a longer-form required more time (in the *combined* and *time* conditions) – an elegant operationalization of speaker production cost. Listener cost was operationalized via a pressure for the matcher to correctly identify the referent.

The authors hypothesized that a preference to map the short form to the frequent object, but not the infrequent object, should emerge in the *combined* condition. Recall that in this condition there was both a pressure to *communicate quickly* (subjects were incentivized based on their time) and accurately (subjects were incentivized based on their matching performance). By contrast, in the *time* condition, in which there was no pressure to communicate accurately (no listener

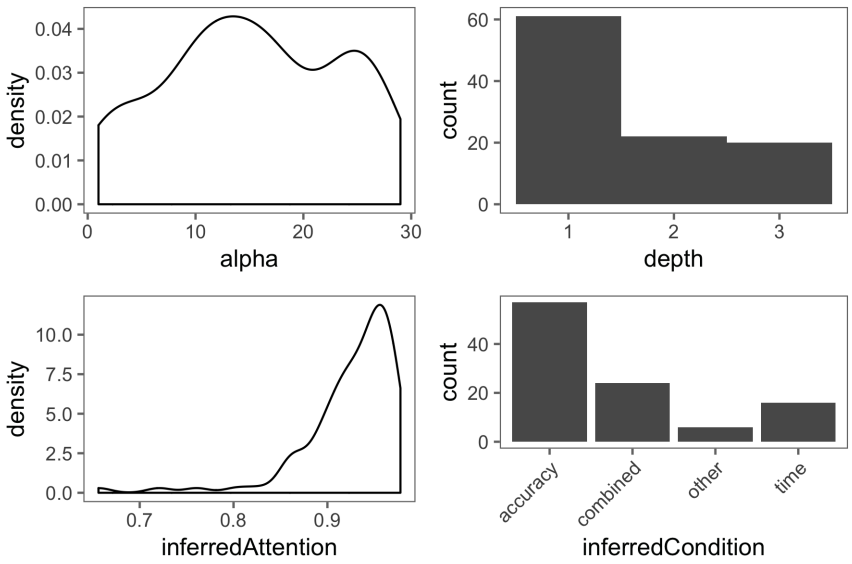

Figure 1: TEMP - NEEDS UPDATING Inferred subject-level parameter distributions. Moving clock-wise from the upper-right facet – inferred alpha values for subjects, the recursive depth for each subject, the probability that subjects are paying attention during the experiment, the inferred cost / referent frequency structure used by the subject.

was present), and only the speaker cost was present, participants should always prefer the shortened form, regardless of the referent. In the *accuracy* condition there was no cost difference between the shortened and lengthened forms, but a listener was present. In this condition, participants should always prefer to use the unambiguous, long-form labels to maximize matching performance. Results across all three conditions were largely consistent with these predictions, but with clear variance in subject-specific strategies by condition. Importantly, only in the *combined* speaker-listener pressure condition did the LOA emerge in many participant's lexicons via a preference to map the shorter form to the most frequent object, (and not the infrequent object).

A changing lexicon or pragmatic language use?

While these results are consistent with a Zipfian account of the LOA, Kanwal et al. (2017) highlight an alternative explanation. They describe a distinction between the "mental representation" of the lexicon and produced forms. In their own words,

"There is a distinction between a language-user's mental representation of the lexicon, and the form-meaning mappings they actually produce in communication... the nature of the communication task in this experiment may have caused them to produce only the short form for one object and the long form for the other based on purely pragmatic considerations."

Put differently, Zipf's Law of Abbreviation is an account of language structure, not necessarily language-use. If participants are retaining both the long-form and short-form words in their "mental lexicon," but *using the lexicon pragmatically*, then the lexicon itself has not "changed." This analysis falls largely in line with Gricean notion's of natural language pragmatics.

Grice and rational, pragmatic language-use

Grice (1975) argued that communication could largely be understood as an instance of rational, cooperative behavior. Under this framework speakers choose utterances to convey meanings. Listeners attempt to infer the speaker's intended meaning given utterances. At the heart of Grice's theory was a set of conversational maxims known to both speakers and listeners (be truthful, relevant, informative and perspicuous). Inferences about a speaker's intended meaning could be derived from speaker behavior relative to these maxims. In particular, this framework assumes that the listener can reason about the speaker's intentions and relevant contextual information (e.g. common ground).

Translating this framing to the current project, we assume that the "matcher" can use contextual information. For example, in the *combined* condition the matcher knows that it is mutually beneficial for them to complete the task as quickly and as accurately as possible. They know the short label is faster (less costly to the speaker) than using a long-form, and that the frequent object occurs three times more often than the infrequent object. In a given trial, in which the matcher is trying to complete the task, a Gricean description of their reasoning might go as follows:

If the director is trying to transmit information as quickly as possible they should always use the faster, shortened form. But if they want me to pick out the correct referent they

should always transmit the unambiguous, slower form. An optimal trade-off should use the short form in as many trials as possible without incurring undo ambiguity. Therefore we should map the shorter, ambiguous form to the more frequent object.

Notably, by reasoning about one-another, about the goals of the game (be quick, be accurate), and the details of context (differential word costs and object frequencies) a pair can converge on optimal *use* of the provided language *without changing the underlying structure*.

Rational speech act theory as a model for pragmatic participant behavior

To model this type of rational, pragmatic interaction we adopt the Rational Speech Act framework (RSA) (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA is a recursive Bayesian model of pragmatic language use, which can largely be seen as a mathematical formalization of essential Gricean principles. RSA has proven to be a productive framework for modeling a range of pragmatic phenomena in both language production and language use including hyperbole, metaphor, implicature and others (CITATIONS).

In the RSA framework, a "speaker agent" defines a conditional distribution, mapping meanings $u \in U$ to utterances $m \in M$, written as $S(u|m)$. We consider a prior over utterances $P(U)$ as well as a prior over meanings $P(M)$. A "listener agent" defines a conditional distribution mapping from utterances to meanings, written as $L(m|u)$. To capture recursive reasoning between interlocutors, each of these functions is described in terms of the other. That is,

$$S_i(u|m) \propto e^{-\alpha \times U(u;m)} \tag{1}$$

where

$$U(u;m) = -log(L_{i-1}(m|u)) - cost(u) \tag{2}$$

and

$$L_{i-1}(m|u) \propto S_{i-1}(u|m) \times p(m) \tag{3}$$

Defining nested speaker and listener agents could, in principle, lead to infinite regress. RSA defines a *literal listener*, denoted $L_0(m|u)$, as a base-case. The literal listener does not reason about a speaker model, rather this agent considers the literal semantics of the utterance.

$$L_0(m|u) \propto \delta_u(m) \times p(m) \tag{4}$$

with

$$\delta_u(m) = \begin{cases} 1, & \text{if } m \in [[u]] \\ 0, & o.w. \end{cases} \tag{5}$$

Bayesian data analysis with RSA linking function

To assess the degree to which the results in Kanwal et al. (2017) can be described by a pragmatic, language-use account, we conduct a Bayesian data analysis, using RSA to model subject data. Our analysis proceeds in two parts. In the *parameter inference* phase we infer subject-level parameters. In the *model-checking and posterior prediction* phase we assess how well our model, with the inferred parameters, predicts the actual production data.

Parameter inference

We model participants as RSA agents, inferring for each the following set of parameters – α and β are RSA model parameters corresponding the the subjects rationality and recursive depth, respectively. The parameters λ and γ correspond to experiment-specific parameters. We describe each parameter in more depth here,

Rationality (α) RSA frames natural language understanding and generation as probabilistic inference. At the heart of this model is a utility calculation (equation 2) in which the speaker considers both the informativity of using an utterance u to convey some meaning m, as well as its cost ($cost(u)$). We capture the degree to which the speaker has as preference to choose the most informative utterance by multiplying the utility function by a scalar α. We infer a separate α value for each subject using a uniform prior over the integers 1...30.

Recursive depth (β) Perhaps most characteristic of RSA is the nested model definitions in which speakers and listeners are defined recursively in terms of one another. The parameter β controls the depth of recursion. We assume a uniform prior over the integers 1..3 inferring the level of recursion for each subject.

Cost and meaning structure (λ) Rather than explicitly map each subject to one of the three conditions of we consider in this project (*combined*, *time*, and *accuracy*) we allow the model to infer these from the data. For each condition we describe the particular utterance cost and meanings prior structure. In the *combined-strategy* condition we consider *utterance costs* in which longer forms are twice as expensive as shorter forms and *meaning priors* reflect actual training proportions from Kanwal et al.(2017) (the infrequent item occurs 8 times and the frequent item 24 times). In the *time-strategy* condition we assume the same structure as the *combined-strategy* condition except we only consider an S_0 speaker who does not consider a listener agent. (Recall that no listener agent is present in the experimental *time* condition.) For the *accuracy-strategy* we consider the same structure as in the *combined-strategy* condition, except we assume that there is no cost difference between the utterances. We also consider a fourth condition *other-strategy* condition, which allows us to describe distintinctly suboptimal production behavior. Some

respondents nearly always used the shorter form for the infrequent object. This behavior is consistent with having reversed the object prior frequencies. We use a categorical prior over these cost-meaning structure assuming that the the *other-strategy* condition is far less likely than the others, equivalent to a Dirichlet prior with alpha values $\alpha = [10, 10, 10, 1]$.

Attention during the experiment (γ) Finally, we consider that subjects may be employing a rational strategy, but might be distracted during the experiment. Our attention parameter γ captures the probability that a participant chooses an utterance randomly in a given trial. We use a beta prior $Beta(a = 12, b = 1)$ prior to encode the expectation that participants are likely to pay attention and are unlikely to choose randomly during the experiment.

Inferred parameter values We approximate the posterior distribution over parameter values with $n = 200$ samples for each subject, using the Metropolis-Hasting algorithm (CITATION). Figure 1 plots the distribution of inferred parameters across all participants. The top-left facet displays the inferred alpha values, which appear largely diffuse across participants. The top-right facet plots recursive depth, showing that most participants are identified as having just a single level of recursion. The bottom-left facet plots the inferred attention (γ) indicating that the vast majority of participants were assigned high-attention values. The bottom-right facet plots the inferred cost and meaning structures. Note that we infer that many participants behaved in a way that was consistent with no cost difference between the short- and long-forms (*accuracy-strategy*) and very few behaved in the strictly sub-optimal *other-strategy*. We believe this particular distribution is consistent with the actual variance in strategies employed by participants in Kanwal et al. (2017), many of whom displayed production strategies that were not strictly condition-optimal.

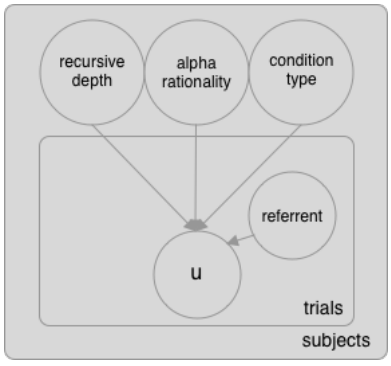

Figure 2: THIS IS JUST A PLACE-HOLDER

Posterior prediction and model checking The posterior predictive distribution describes the data we should *expect to see* sampling from our fitted model. Intuitively if these predictions do not match the *actual* data used to fit the model, then we have a poor model. Recall that in the first step of our Bayesian data analysis we infer parameters for each subject including the rationality parameter (α), their recursive depth (β), the cost and referent condition structure (λ), and their level of attention (γ).

Having inferred subject-level parameters we would like to "run" our RSA-subjects through the original experiment. Operationally, this means that we consider the actual sequence of trials for each human subject, generating a corresponding sequence of utterances for each of the trials they saw. For example we consider the j^{th} trial for subject i. We pass the actual object referent subject i observed in the jth trial (m_{ij}) to our RSA instantiation of that subject and generate an utterance (u_{ij}). This amounts to sampling an utterance $u_{ij} \sim S(u_{ij}|m_{ij};\theta_i)$ where θ_i contains the particular parameters inferred for subject i (i.e. $\theta_i = \{\alpha_i, \beta_i, \lambda_i, \gamma_i\}$). Having "run" each of our RSA subject-representations through the experiment we can compare model production data to actual Kanwal et al. (2017) participant production data.

Results

Following the analysis in Kanwal et al. (2017) we examine the proportion of trials in which a particular subject used the short-form (see Kanwal Figure 3). That is, we compare the proportion of trials in which subject i used the short form, comparing this to proportion of trials predicted by our RSA model and by a baseline comparison.

Baseline model As a baseline we compare a deterministic speaker using the optimal Zipfian language in each condition. Under this model, a speaker in the *combined* condition should always map the short form to the most frequent item, but not the infrequent item. In the *time* condition, the speaker should always use the short form, regardless of the referent. In the *accuracy* condition, the speaker should always use the long form. This baseline describes a speaker who deterministically chooses an utterance based on the optimal strategy.

To assess model performance we correlate model posterior predictive values for our RSA- and baseline-models to human data. Table 2 displays r^2 values for both models. We find close fit for our RSA agents, significantly above our baseline model (NEED STAT TEST), accounting for a large proportion of participant-level variance. Figure 3 displays model fit between baseline and BDA-RSA predictions to human data.

Model	r^2
Baseline	0.466
BDA RSA	0.972

General discussion

Zipf (1935) argued that human behavior could largely be understood in terms of effort minimization. In the domain of language, he argued that the particular distributional properties we find in large-scale corpora, such as the fact that more frequent words tend to be shorter, provide evidence for his effort-minimization framework. Kanwal et al. (2017) presented an experimental test of the Zipfian hypothesis, conducting a miniature language-learning experiment and isolat-

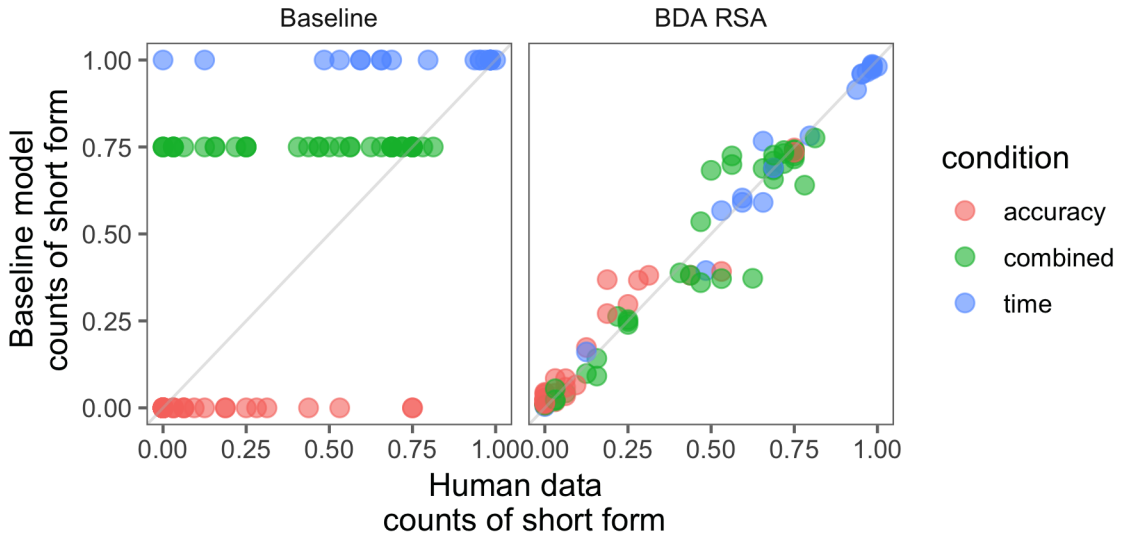

Figure 3: Posterior predictive values lead to significantly better fit to human data than a baseline model. Horizontal axes display proportion of short-form usage from subjects in Kanwal et al. (2017). Vertical axes display model predictions. The left facet displays predicted short-form usage under a baseline model that implements the 'optimal Zipfian language for each condition. The right facet displays predicted short-form usage under our BDA-RSA model.

ing the impact of speaker- and listener-pressures directly. In their analysis Kanwal et al. (2017) highlighted that their results were consistent with a Zipfian, language-change account as well as a pragmatic language-use account. This distinction, they argued, cannot be assessed with the current set of results as the data collected only included production behavior (and no information about actual, underlying lexicons being used by subjects).

To test the pragmatic, language-use account we conducted a Bayesian data analysis, modeling subjects as rational pragmatic agents using the Rational Speech Act framework (Frank & Goodman, 2012; Goodman & Frank, 2016). Inferring subject-level parameters we find good fit with the empirical data providing evidence that subject behavior in this experiment may be a result of pragmatic language use, rather than lexicon change.

Limitations of the current analysis While our current framework accounts for a large percentage of variance in the human data, there are aspects our current analysis does not capture. Figure 5. of Kanwal et al. (2017) shows temporal effects throughout the experiment – speakers are more likely to use the shorter form later in the experiment compared to earlier in the experiment in the *combined* condition. Our current model does not capture this particular effect. However, this effect could be due to a variety of factors modeleable in our current framework. One such perspective might relate to uncertainty over object frequency priors diminishing as the experiment progresses. In the current framework, a preference to use the short-form in the *combined* condition is modulated by the relative proportion of the frequent and infrequent object – holding all else constant, the greater the relative difference in object frequencies, the greater the preference to use the short-form. If participants were uncertain about the prior frequencies (or that they shared the same priors with their interlocutor) but, became more confident over trials that the more frequent object would appear in the same proportion then we should their preference to use the short-form for that object increase.

A separate approach might consider the temporal effect as emerging from increased sensitivity to production cost as the task proceeds. Essentially, waiting for the message to transmit may become more onerous over time, making speakers more likely to want to avoid it. Our current framework provides ample opportunity to investigate these and related hypotheses about the exact nature of participant behavior.

The connection between rational language use and language change Recent theoretical and computational work has highlighted connections between local, in-the-moment language use as modeled by rational, pragmatic agents and aggregate, distribution-level properties of natural language (Goodman & Frank, 2016; Levy, 2018). This is not without precedent. Zipf (1935) implicitly argues language may be optimized at the level of conversation in his example of an optimal speaker and listener language. He describes a set-up in which there are a set of utterance U and meanings M which differ in costs and frequency, respectively. While Zipf did not explicitly described the level at which language optimization occurs, his example mirrors the reference game setting we have adopted here.

We note that the connection between Zipfian principles

and Gricean, conversational pragmatics was first described by Horn (1984), who argued that Zipf's speaker and listener pressures ("Speaker's economy" and "Auditor's economy") were fundamental to Grice's (1957) derivation of his Cooperative Principle and associated conversational maxims. In general, it is not hard to imagine a scenario in which pragmatic language use bootstraps later lexicon-level change via processes like intergenerational transmission. Kanwal et al. (2017) present this very idea writing,"[whether] pragmatics-driven asymmetry in usage may or may not lead to an immediate shift in lexical representations, it may be an important first step in such a change." Take the example of the current experiment. In the *combined* condition some subjects produced languages which only consisted of the shortened form for the frequent object and the long form for the infrequent object. If a subsequent generation of speakers were to learn from this data, they would never be exposed to the original mapping from the frequent object's original long-form. In this way pragmatics may serve as an essential precursor to more system-level language change.

Acknowledgements

Fill in

References