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

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 languages emerging from the pressure to minimize speaker-listener effort. Subsequent work has provided indirect evidence for a Zipfian analysis of this relation between frequency and length, what Zipf termed the "Law of Abbreviation". In an effort to directly isolate the impact of speaker and listener pressures, Kanwal et al. (2017) conducted a study using an artificial language learning paradigm. The authors showed that only when both a speaker and listener pressure were present did languages consistent with the Law of Abbreviation emerge. The authors highlight that while their data is consistent with a Zipfian analysis, it is also consistent with pragmatic langage use account. We consider this latter account conducting a Bayesian data analysis using the data from Kanwal et al. (2017) representing subjects as rational, pragmatic agents using the Rational Speech Act framework. Fitting subject-level parameters we find good fit between model posterior-predictive values and data from Kanwal et al. (2017). We argue this analysis provides evidence for a connection between Zipfian notions of efficiency driven language change and Gricean notions of rational-pragmatic langauge use.

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

Zipf (1935) presented a view of human behavior that focused on *effort minimization* as a guiding principle. Relying on empirical evidence from 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 later termed the "Law of Abbreviation," Zipf claimed that the more frequent words are, the shorter they tend to be. Evidence for this relationship can be found across languages, may also be present in animal communication systems (Ferrer-i-Cancho et, 2013), and is a property of artificial (programming) languages as well.

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. Importantly, there remains a fundamental asymmetry in any communicative act – what is effortful for a speaker (production) is different than what is effortful for a listener (understanding). Zipf suggested that an optimal language from a speaker perspective would consist of a single, low-cost word which was fully ambiguous - referring to any possible referent in the world. In using 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 could also be explained by a process of random typing (Ferrer-i-Cancho & Mascoso del Pardo, 2012). Moreover, most evidence supporting the Zipfian approach has been indirect and did not fully determine the causal role of competing speaker and listener pressures.

Kanwal et al. (2017) implemented an artificial language learning paradigm to isolate speaker-listener 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 "Law of Abbreviation" emerge. The authors adopted a reference game setting in which participants learned a small language consisting of three words, which could be used to refer to two items. They hypothesized that when given a choice between words subjects should prefer to use the less costly, but ambiguous term to refer to a more frequently occurring object. Importantly, however, this choice should only appear when there was both a speaker- and listener-pressures present. Results largely confirmed their hypotheses.

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 "Law of Abbreviation" is a lexicon-level theory. That is, it attempts to explain why particular word forms are mapped to particular meanings. However, their results are, in theory, consistent with a pragmatic language-use account.

"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 consider-

ations."

Put differently, 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' correctly point out that because they only recorded participants' actual language production they cannot distinguishing between the lexicon change account and pragmatic accounts.

The current project addresses the question of whether the 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 ($R^2 = 0.986$), over and above what would would be expected from a baseline model ($r^2 = 0.68$). 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 pragmatic language use. We evaluate our results summarising the findings and highlighting aspects that our modeling does not 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. We end with a summary of our findings and next steps.

Experimental evidence of Zipf's Law of Abbreviation

Kanwal et al. (2017) used an artificial language learning paradigm to isolate the impact of speaker- and listener-pressures on language structure. During training, participants first learn a miniature language. During test, participant's language use is recorded, typically over a series of communication events or across multiple generations of participants. This paradigm has been used to study changes in language structure (CITATIONS).

Kanwal et al. (2017) adopt this basic framework – participants learned a miniature artificial language and were 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. The study proceeded as follows. Participants were assigned to one of four possible conditions. A *combined* condition in which both speaker- and listener-pressures were present. An *accuracy* condition, in which only a listener pressure was present. A *time* condition in which only a speaker pressure was present. They also tested a *neither* condition in which neither speaker nor listener pressures were present. For the purpose of our current project, we do not model the *neither* condition as it's unclear what optimal behavior should like.

For a comprehensive description of the experimental set-up we refer the reader to Kanwal et al. (2017), focusing on essential details in this section. During training participants learned labels for two novel objects. Importantly, one of the objects occurred more frequently than the other. 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, pairs of participants were asked to participate in a communication game in which a "director" was shown an object" and asked to transmit its name 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 – an elegant operationalization of speaker production cost. Listener pressure 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 more frequent object, but not the infrequent object should only occur in *combined* condition. By contrast, when there was no pressure to communicate (no listener was present in the *time* condition), and only the speaker cost was present, participants should always prefer the shortened form, regardless of the referent. When there was no cost difference between the shortened and lengthened forms, but a listener was present, then participants should always prefer to use the unambiguous, long-forms regardless of referentin the *accuracy* condition. Results were largely consistent with these hypotheses.

A changing lexicon or pragmatic language use?

While the data is consistent with a Zipfian account of the Law of Abbreviation, Kanwal et al. (2017) highlight an important alternative explanation. They describe a distinction between the "mental representation" of the lexicon and produced forms. That is, Zipf's Law of Abbreviation is an account of lexical change, not language-use. If participants are retaining both words in their "mental lexicon," but *using the lexicon pragmatically* by using the shortened form only for the more frequent item, then the lexicon itself has not "changed" at all. This analysis falls largely in line with Gricean notion of natural language pragmatics.

Grice (1957) argued that communication could largely be analysed as an instance of rational, cooperative behavior. Under this framework speakers choose utterances to convey meanings. Listeners attempt to infer the speaker's communicative goal. 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 these maxims by a listener. In particular, it is assumed that the listener can reason about the speaker and information that is mutually known to both. The matcher can use such information, such as the fact that using the short-form is less costly than using a long-form, and that the frequent object occurs three times more often than the infrequent object. The matcher and director might reason as follows: *If the director is trying to transmit information as quickly as possible they should always use the quicker, shortened form. But if they want me to pick out the correct referent they should always transmit the slower, longer form. An optimal solution should use the short form in in as many trials as possible. 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, but *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 (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA is a recursive Bayesian model of pragmatic language production and interpretation, 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 from hyperbole, metaphor, scalar implicature and others (CITATIONS). Under RSA a speaker agent describes a conditional distribution mapping meanings $u \in U$ to utterances $m \in M$, written as $S(u|m)$. A speaker agent describes a conditional distribution mapping from utterances to meaning, written as $L(m|u)$. Importantly, 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 recursion. We take 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.

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

Where

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

Bayesian data analysis with RSA linking function

Top assess the degree to which the observed production data in Kanwal et al. (2017) can be described by a pragmatic language-use account we conduct a Bayesian data analysis, using RSA as our linking function. 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, describes the experimental data.

Parameter inference For each participant we infer a set of parameters – α, β are RSA model parameters corresponding the the subjects rationality and recursive depth, respectively. The parameters γ and λ correspond to the particular condition (made up of utterance costs and object referent priors) as wells as a focus parameter describing the probability that the subject is not paying attention and choosing to produce utterance randomly. Table 1 includes the parameters and their descriptions.

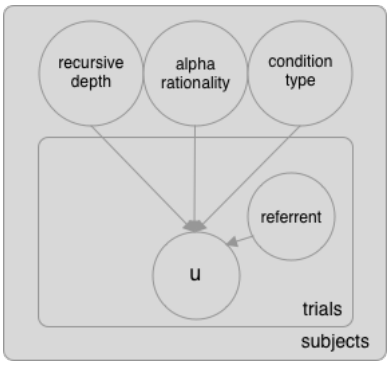

Figure 2: THIS IS A PLACE-HOLD AND IS INCORRECT

For each subject we use $n = 200$ samples using Metropolis-Hasting to approximate the posterior distribution over parameters. Figure X displays inferred subject params.

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 parameters for each subject including *alpha* the rationality parameter, γ their recursive depth, λ

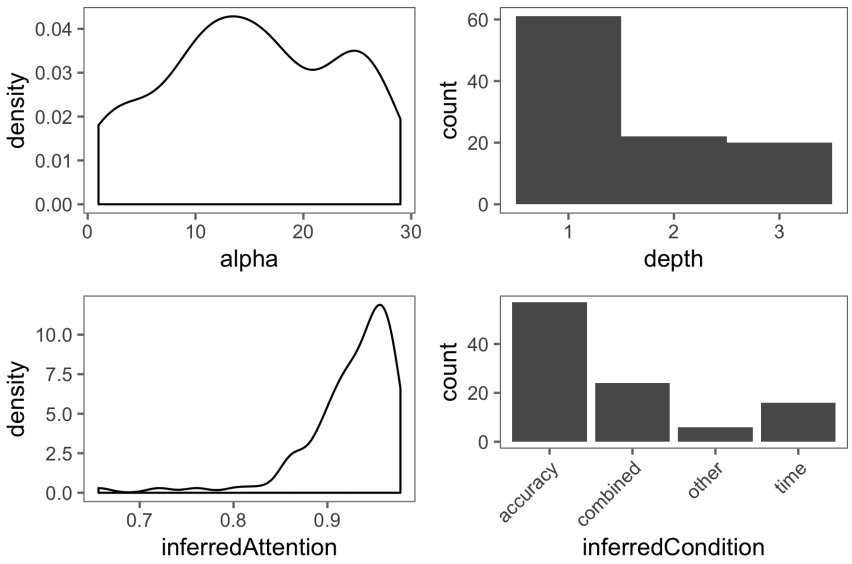

Figure 1: TEMP - NEEDS UPDATING Inferred subject-level parameter distributions. Moving clock-wise from the upper-right facet – inferred alpha values for subjects, the recursive depth for each subject, the probability that subjects are paying attention during the experiment, the inferred cost / referrent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

the cost and referent condition and finally σ, the degree to which the participant was paying attention during the experiment. Figure X displays these parameters.

Having inferred these subject-level parameters we'd like to "run" our RSA-participants back through the experiment. That is, for each participant and for each trial they saw (including the object they needed to refer to in that trial) we sample an utterance. For example, for a participant who was in the *combined* condition we would sample an utterance given that the actual participant viewed some object on trial 1, repeating the same process for all 32 trials, 24 of which would require the model to refer to a frequent object and 8 requiring the model to refer to an infrequent object. Having simulated datasets for each participant we can compare the number of times the RSA-agent and actual participant used the short-form in each condition. Correlating model posterior predictive values to human data find close fit $r^2 = 0.987$. Figure X compares these values.

As a baseline we can compare an "optimal Zipfian model". Under this model a speaker in the *combined* condition should always map the short form 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 forms. For comparison, model fit under this model achieves an overall $r^2 = 0.630$.

Results

Limitations of the current analysis There are aspects our current analysis does not capture – the temporal effects identified by the authors – however, this remains to be a possible extension of the framework we introduce here. Moreover we frame our results in light argue that this is largely in line with ideas put forth by the authors – that pragmatic language use may be an essential precursor to more system-level language change. In their own words, "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."

Connecting pragmatic language use to language change

Conclusion

Languages around the world display a consistent relationship between word length and frequency – more frequent words tend to be shorter (CITATIONS). Zipf (1935) called this particular finding the "Law of Abbreviation" and characterized it in terms of competing speaker listener pressures. While other work has provided indirect evidence consistent with this Zipfian analysis (CITATIONS), Kanwal et al. (2017) attempted to derive this effect while isolating speaker and listener pressures experimentally. In an artificial language learning paradigm the authors showed that subjects were significantly more likely to produce lexicons consistent the "Law of Abbreviation," but only when both speaker and listener pressures are present. While this evidence is compelling the authors highlight an alternative hypothesis that may have generated the same effect, but could not be distinguished in the current study – pragmatic language use instead of lexical change. To test this secondary hypothesis we conducted a Bayesian data analysis, modeling participants as rational pragmatic agents using the Rational Speech-act (RSA) framework. Inferring subject-level parameters we are able to fit

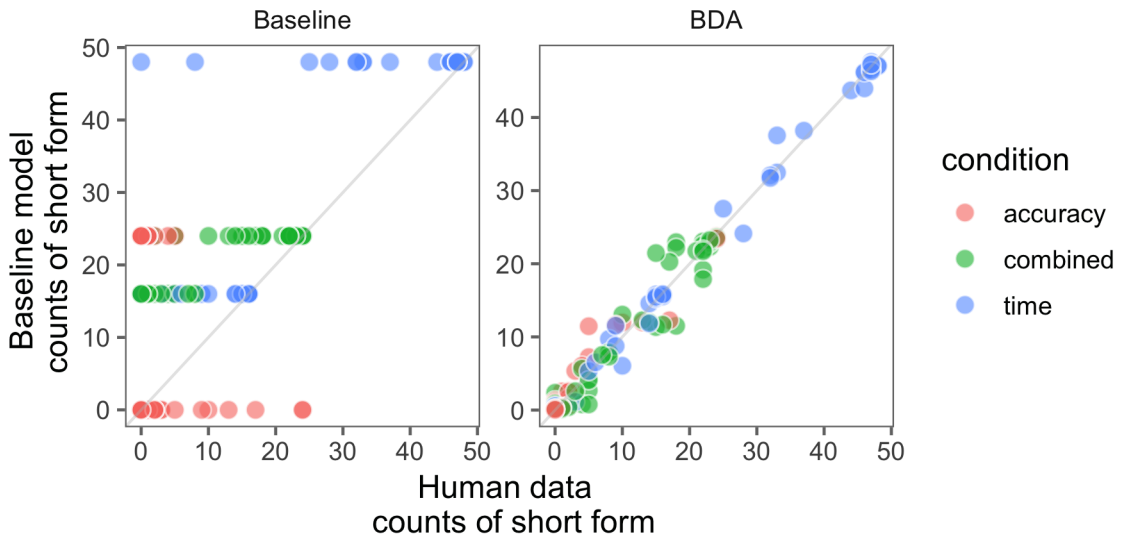

Figure 3: Posterior predictive values lead to far better fit to human data than a baseline model. Horizontal axes display counts of short-form usage from subjects in Kanwal et al. (2017). Verticle axes display model predictions. The left facet displays predicted short-form usage under a baseline model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

Kanwal et al. (2017) experimental data with high-levels of accuracy ($r^2 = 0.986$). While we believe this provides evidence for the pragmatic-language use interpretation we also highlight the compatibility of this interpretation with an additional insight of Kanwal et al. (2017) – pragmatic language use may serve as an important bootstrap for future language change.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References