Partial truths: Adults choose to mention agents and patients in proportion to informativity, even if it doesn't fully disambiguate the message

Melissa Kline

Laura Schulz

Edward Gibson

Abstract

How do we decide what to say to ensure our meanings will be understood? The Rational Speech Act model (RSA, Frank & Goodman, 2012) asserts that speakers plan what to say by comparing the informativity of words in a particular context. We present the first example of an RSA model of sentence level (who-did-what-to-whom) meanings. In these contexts, the set of possible messages must be abstracted from entities in common ground (people and objects) to possible events (Jane eats the apple, Marco peels the banana), with each word contributing unique semantic content. How do speakers accomplish the transformation from context to compositional, informative messages? In a communication game, participants described transitive events (e.g. Jane pets the dog), with only two words, in contexts where two words either were or were not enough to uniquely identify an event. Adults chose utterances matching the predictions of the RSA even when there was no possible fully 'successful' utterance. Thus we show that adults' communicative behavior can be described by a model that accommodates informativity in context, beyond the set of possible entities in common ground. This study provides the first evidence that adults' language production is affected, at the level of argument structure, by the graded informativity of possible utterances in context, and suggests that full-blown natural speech may result from speakers who model and adapt to the listener's needs.

Introduction

Communication requires continually making decisions about what information to include and exclude. It is not always necessary to fully describe an event: if someone asks *What are you doing*? then *I'm eating* might be sufficient, and possibly preferable to longer alternatives like *I'm eating a sandwich* or *I'm eating a grilled cheese sandwich*. For a speaker to successfully communicate with a listener in this way, the two need to implicitly agree on some shared principles of communication. Grice (1975) codified these conversational assumptions as a series of 'maxims', including the maxims of Quantity ('give as much information as is needed, but no more') and Relevance ('say something that furthers the goal of the conversation'). Thus a speaker can refer to *a sandwich* alone if the alternative is a salad, but should refer to *a grilled cheese sandwich* if the alternative is peanut butter and jelly.

As listeners, adults understand language in part by using statistical information to predict upcoming words and structures (Altmann & Kamide, 1999; Levy, 2008; MacDonald, 2013; MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; cf. Kuperberg & Jaeger, 2016 for a recent review.) How does this predicting listener operate? Listeners could simply expect similar language to occur in similar contexts, without regard to the speaker's motives. But more specifically, they could expect speakers to behave predictably because they expect them to behave *helpfully*. A recent formalization of this latter hypothesis is the Rational Speech Act model (RSA), which is based around a cooperative speaker-listener pair (Frank & Goodman, 2012). Speakers attempt to maximize the information transferred to the listener, and listeners succeed by assuming that the speaker is doing this. RSA models successfully predict a variety of phenomena in pragmatics including the interpretation of scalar implicatures, hyperbole, and metaphor (Goodman & Frank, 2016; Goodman &

Stuhlmüller, 2013; Kao, Levy, & Goodman, 2013; Kao, Wu, Bergen, & Goodman, 2014). Are listeners warranted in making these generous assumptions about speakers? Many features of language production seem to be shaped to improve the chances of successful communication. Formal approaches based on information theory (Shannon, 1949) have been used to successfully explain reduction and omission phenomena in natural language production including phonological reduction, lexical choice (e.g., *math*/*mathematics*) and inclusion of optional arguments (Aylett & Turk, 2004; Jaeger, 2010; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Resnik, 1996; van Son & van Santen, 2005; though see Keysar, Barr, & Horton, 1998).

If production is driven by the value to the listener rather than the costs to the speaker, then the speaker should flexibly adapt when the (linguistic *or* non-linguistic) context changes. For the specific case of referring expressions (*that, that big sandwich*), there is a large body of work showing that speakers' choices are related to available nonlinguistic information (e.g. Brennan & Clark, 1996; Brown-Schmidt & Tanenhaus, 2008; Nadig & Sedivy, 2002; Pogue, Kurumada, & Tanenhaus, 2016; Sedivy, 1999). This is taken as evidence of an awareness of listeners' needs because the language production cost of *that big sandwich* is presumably the same across contexts, while the benefit to the listener is considerable when there are many sandwiches, but null if the listener can already pick out the lone sandwich. Speakers do this even when a listener would need to make inferences about a speaker's intention to succeed: in a context with a *blue circle* as a target with a *blue square* and a *green square* as distractors, adults limited to a single word produce CIRCLE to identify the target object, not BLUE: although BLUE is a good description of the target in isolation, it could also refer to the blue square (Frank & Goodman, 2012).

But human language goes beyond referring expressions for objects: sentences express entire propositions about the world (*Ben is eating my grilled cheese sandwich*.) Deriving the set of possible propositions (not just possible object referents) would seem to require an extensive understanding of both world knowledge and the ways that conversations tend to unfold (cf. Ginzburg, 1996). Even once a particular proposition has been chosen, we have many choices about how to encode it in a sentence. We make choices about argument structure and verb identify (*he ate it/he put it in his mouth*), and language provides many ways to omit or limit how much we say in conveying a proposition, including pronouns (*Ben/he ate the sandwich*), ellipsis (*Ben ate the sandwich, and then a cookie*), passive constructions (*The sandwich was eaten*), and optional arguments (*Ben ate the sandwich [with a fork and knife]*). These options are used pragmatically: speakers tend to (a) omit or reduce information the listener can retrieve from linguistic context, (b) converge with dialog partners on syntactic alternations, and (c) include optional material when listeners might otherwise go astray (Brennan & Hanna, 2009; Galati & Brennan, 2010; Horton, 2005; Kurumada & Jaeger, 2015; Pickering & Garrod, 2004).

Relatively little attention has been paid to how speakers use *nonlinguistic* information to produce informative sentences. Do we attempt to communicate sentence-level meaning using something like the rational speaker model, tailoring what we say to the surrounding context? At least one study suggests this may be the case: Lockridge & Brennan (2002) had participants describe scenes with either typical or atypical instruments (*He stabbed him with a knife/an icepick*) to a naïve listener. In an unconstrained storytelling task, speakers were more likely to mention atypical than typical instruments, especially when the listener could not see the event. However, understanding event descriptions is challenging exactly because events are transitory – they don't 'stick around' in the context like objects do, and references to events often occur when

the event itself is in the past or future (Gleitman, 1990). Thus, while this study suggests speakers are sensitive to how world knowledge impacts linguistic informativity, it does not address the fit between production and particular non-linguistic contexts: the contrast in that study is between not seeing the event (the usual scenario) and seeing the event as it is being described (which listeners usually can't.)

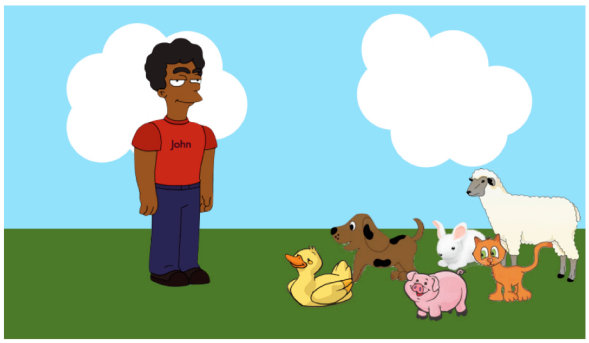
To begin understanding how speakers use nonlinguistic context to decide what to say about an event, we focus on a single class of basic propositions: transitive sentences like *John feeds the dog*. While this construction is used for many classes of verbs, we consider prototypical cases in which the basic meaning involves an *agent* performing some action on a *patient*. Even assuming the speaker is referring to an event that might occur (or has recently occurred) in the immediate context, the set of possible messages is potentially infinite (Quine, 1960). Here, we leave aside the question of word-to-concept mapping and focus on the question of possible events given a set of possible participants. A speaker trying to design an informative event utterance must consider not only the possible verbs, but also what referent could correspond to each argument position. We can represent the number of possible events as the product of the possible verbs, agents, and patients:

(1) {John, Sue, George, Maria, Jenny} x {feeds, chases, pets} x {the dog, the cat} =30 events

We use this logic to create 'toy' worlds in which there are always exactly seven entities (people and objects), and the messages to be communicated are interactions between these entities (e.g. *John feeds the dog*).

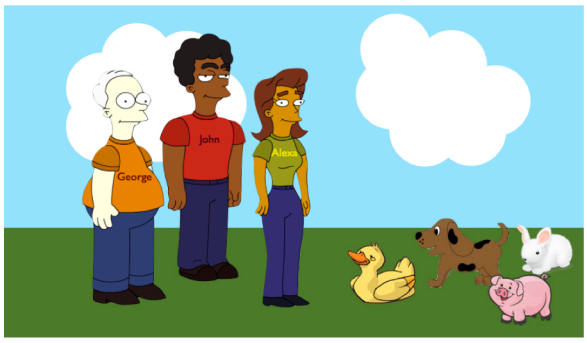
In natural speech, both agents and patients can sometimes be omitted from transitive descriptions. Many transitive verbs can be used intransitively, e.g. *We'll eat in the kitchen*[1]*,* and many languages also allow noun phrases in subject position to be omitted relatively freely (e.g. in Spanish *Comió bocadillos, (He) ate sandwiches*.) In English, these kinds of subject omissions require specific discourse context (e.g. a command, *Don't eat in the kitchen*). We therefore use a production task that restricts the producer to exactly two words, forcing participants to make the choice to omit at least one element (agent, patient or verb). In most object reference studies (cf. Brennan & Clark, 1996; Brown-Schmidt & Tanenhaus, 2008; Nadig & Sedivy, 2002; Pogue, Kurumada, & Tanenhaus, 2016; Sedivy, 1999), a noun phrase like *my sandwich* or *my grilled-cheese sandwich* is assumed to be informative when it uniquely identifies one out of several referents in the context, under-informative if it could apply to more than one object (two such sandwiches), and over-informative if it includes additional modifiers (*my grilled cheese sandwich* when there is only one sandwich). RSA models assume a richer sense of 'informativity' in which words are informative to the extent that they reduce the number of possible interpretations by any amount (Frank & Goodman, 2012). Thus, we can vary the informativity of these utterances by varying the possible events that might have occurred in the local context, specifically by manipulating the set of possible agents and patients. We can then ask whether speakers choose informative utterances, even in cases where a listener would be unable to identify the entire event meaning.

[1] These unergative alternations can be contrasted with unaccusative intransitive alternations like *John broke the lamp*/*The lamp broke*; here we focus solely on the inclusion or omission of agents and patients rather than on the argument structure or syntactic behavior of particular verbs.



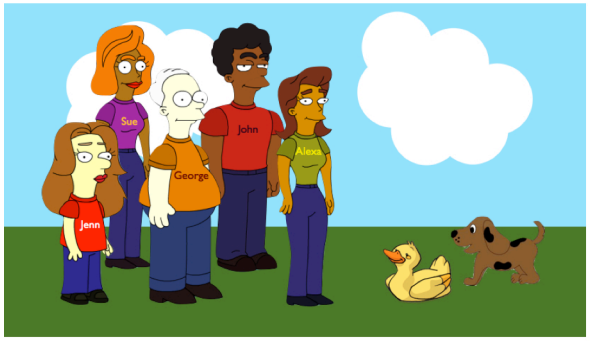
(1a) 1 agent & 6 patients, [1:6]

(1b) 2 agents & 5 patients, [2:5]

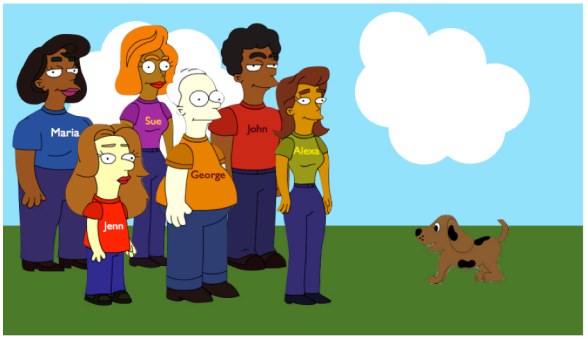


(1c) 3 agents & 4 patients, [3:4]

(1d) 4 agents & 3 patients, [4:3]



(1e) 5 agents & 2 patients, [5:2]



(1f) 6 agents & 1 patient, [6:1]

(1g) Target event

Figure 1: Context and event images for JOHN FEEDS THE DOG. We refer to each context condition by the number of people (agents) and objects (patients) present, e.g. Fig. 1a is notated as [1:6], Fig 1b as [5:2], and so on.

Figure 1 shows a possible event (JOHN FEEDS THE DOG) and six sets of entities that could participate in the event to be named. Each context set is made up of people (canonical agents) and either animals or inanimate objects (both of which are more likely than humans to appear as patients). Critically, we manipulate the communicative context (and therefore the informativity of potential utterances) by altering the set of seven entities that appear in the context picture. If the context is Figure 1a, the utterance FEED DOG fails to resolve the ambiguity (anyone could have done it); on the other hand, the utterance JOHN FEED specifies the agent and relies on an intelligent listener to identify the unique patient in context. For Figure 1f, the reverse is true: FEED DOG resolves the ambiguity. In the intermediate cases (Figures 1b-1e), there is no two-word utterance that can fully disambiguate the intended meaning: there are multiple options for both agent and patient, and the verb cannot be uniquely inferred from the context images.

Our critical hypothesis has to do with how people will behave in the four intermediate arrays. In these conditions, different words reduce ambiguity to different degrees: in Figure 1e, mentioning John (and the verb) narrows down the possible events to just two alternatives (he feeds the dog or duck) rather than five (somebody feeds the dog). If the RSA model extends to descriptions of argument structure relations, adults should still be able to select informative utterances: when there are more agents than patients, participants should be more likely to mention subjects, even if ambiguity between multiple messages remains. However, if participants use a simpler strategy of determining just whether or not a given utterance successfully conveys the intended event, then they should still choose informative arguments in the deterministic cases, but perform at chance (or otherwise not differentiate the intermediate conditions) when both arguments remain ambiguous.

Methods

Participants 91 English-speaking adults participated on Amazon's Mechanical Turk (AMT). Participants were screened to be located in the United States and self-reporting English as their first language (an additional 21 participants were excluded who did not meet these criteria). No other demographic information was collected. The task took approximately 13 minutes to complete and participants were paid $1.00. This pay rate was based on an anticipated study length of 10 minutes, following the 10¢/minute rule of thumb used for AMT studies in the lab at the time these data were collected. All participants gave informed consent in accordance with the requirements of the Massachusetts Institute of Technology's institutional review board.

Stimuli We created cartoon stimulus sets for each of twelve verbs (*eat, feed, hold, drink, kick, drop, wash, pour, throw, touch, read*, and *roll*). Each set consisted of an action picture and six 'context' pictures showing possible agents and patients who might participate in the event. The people were generated using a character-creation website (Brooks et al., 2007) with distinct features and names on their shirts. The objects were chosen from a category (e.g., various foods) appropriate for each verb. The total number of agents and patients in each context sums to 7, yielding six variations (i.e. [6:1] to [1:6]) for each of the twelve stimulus sets. All stimuli, code, and analyses are available in the Supplemental Materials for this article.

Procedure Stimuli were presented using Python and the EconWillow package (Weel, 2008), accessed through AMT. Participants were told that they were providing descriptions for another (sham) participant. Participants saw the trials in a random order, with two items presented at each context type. On each trial, they saw the context picture for ten seconds, read a sentence describing the action they would see (e.g., "John feeds the dog"), and then saw the

action picture for ten seconds. Finally, the context picture reappeared and participants were given two separate text boxes to enter their description; if they entered more than two words (screened by checking for spaces, e.g., "baby rolls" in one box), they were told to try again. To encourage participants to answer quickly, their response speed in seconds was shown after every trial.

Data Coding

A total of 1092 responses were collected from the 91 participants, 182 responses in each condition. Responses were first checked for minor variations such as capitalization and verb form (e.g., "Eaten" was coded as "eat"). The majority of these responses (84%) consisted of two of the possible three content words in the sentence (e.g. JOHN FEED, FEED DOG, or JOHN DOG). In the remaining responses, participants deviated from these exact lexical items; in these cases we checked if the word used could refer to a unique entity (e.g. *she* in an array with a female agent among only male distractors). A full record of this coding is available in the Supplemental Materials; just 20 responses (1.8%) consisted of two unclassified words and thus were excluded from analyses. Because not every response contained two codable words, we present analyses below for the presence of agents, patients, and verbs in each response.

Results

We code the main effect of interest numerically, representing the key condition of context type in the model as the number of potential agents in the context image (recall that the number of agents and patients in these context images are inversely related, always summing to seven total). The effect of the number of agents vs. patients on whether participants mentioned the

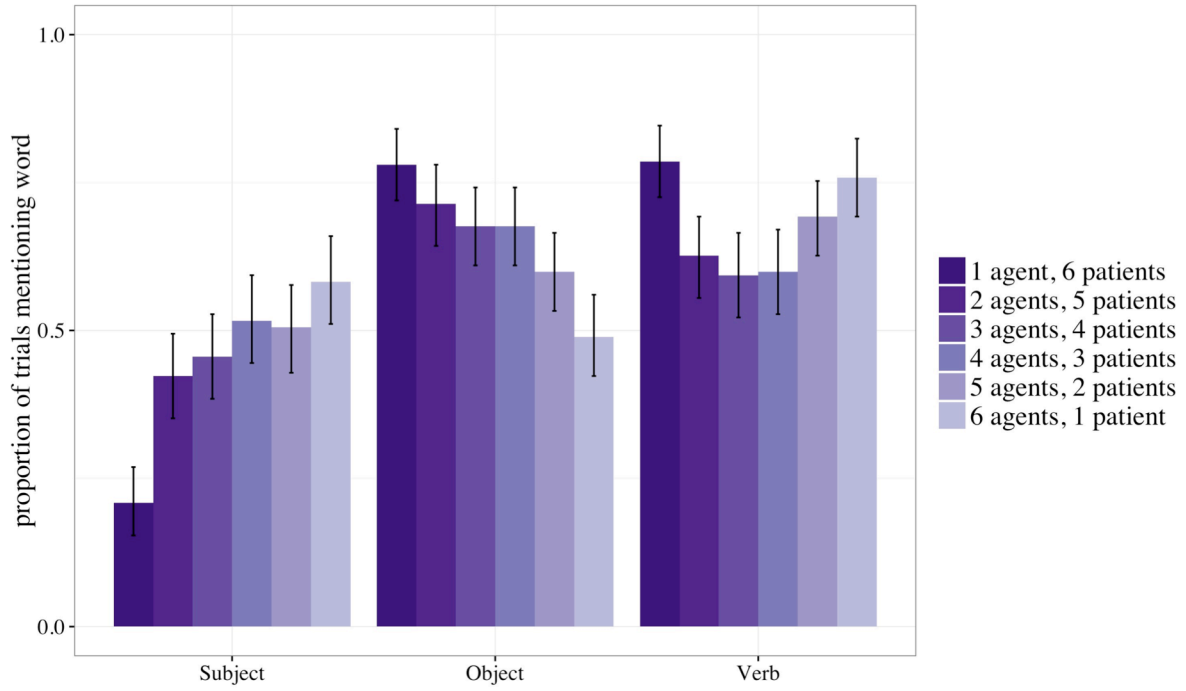


Figure 2: Trials on which participants included subjects, objects, and verbs. Error bars represent 95% bootstrapped confidence intervals.

agent in their response was highly significant by a mixed-effect logistic regression[2] with random slopes and intercepts for both item and participant ($\beta = 0.55$, SE = 0.15, $Z = 3.79$, $p < 0.001$; LRT: $X^2 = 10.7$, df= 8, $p<0.005$). The same was true for patients ($\beta = -0.55$, SE = 0.13, $Z = -4.38$, $p < 0.001$; LRT: $X^2 = 17.2$, df= 8, $p<0.001$). These patterns are as predicted – as more agent distractors (and thus fewer patient distractors) were present, participants were more likely to mention the agent and less likely to mention the patient. We also found that participants overall were somewhat more likely to mention patients than agents: on the subset of trials (74%) where

[2] In addition to reporting beta statistics, we evaluate these models with likelihood ratio tests by comparison with a model with the same random effects and only the effect of interest omitted from the fixed effects structure. Exact model specifications can be found in the analysis file named *MD_turk.R* in the repository for this paper.

participants mentioned only one of the two, there were significantly more patients than agents ($p < 0.001$, binomial test).

To test whether participants gave graded responses to the intermediate arrays (e.g., [2:5]), we also examined the effects of array type after removing trials for which a 'deterministic' answer could be given ([6:1], [1:6].). The effects of array type on both agent and patient mention were both significant when evaluating only these intermediate cases (Agent mention: $\beta = 0.30$, SE = 0.10, Z = 2.97, $p < 0.005$, LRT: $X^2 = 5.89$, df= 8, $p<0.05$; Patient mention: $\beta = -0.33$, SE = 0.16, Z = -2.01, $p < 0.05$, LRT: $X^2 = 3.88$, df= 8, $p<0.05$).

Model Comparisons

To evaluate how human performance might reflect pragmatic choices, we compared three computational models (with two additional variations shown in the Supplemental Materials). Each of these models generates (unordered) two-word utterances ["AV" – agent and verb, "VP" – verb and patient, or "AP" – agent and patient] at each of the conditions in the experiment; we compare model predictions to participants' responses of these types (omitting the ~15% of responses that included some other word). Below, we describe the common assumptions the models share, define the particulars of each model, and then compare them to human performance.

In all models, the shared context is the set of possible events that might occur given the set of agents, patients, and plausible verbs (we assume the prior probability of picking any particular event *e* in *E* is uniform). We assume that each object/person in the scene is classified unambiguously as an agent or patient (wrong in general, but true in our experimental context). For the verbs, we assume participants are considering some set of possible interactions between

the agents and patients (e.g., petting, feeding). In principle, the notion of 'possible verb set' could be estimated empirically by asking naïve participants to list possible actions between the agent-patient stimulus sets directly. Here, we simply assume the set is relatively small and does not vary with the number of agents and patients[3]. Thus, the shared context of possible events **E** is

(2)

$$E = \{(a_1, v_1, p_1), (a_1, v_1, p_2), (a_1, v_2, p_1) \ \ldots\}$$

With the number of possible events *e* in **E** calculated as follows:

(3)

$$A = \{a_1, a_2, \ldots, a_i\}$$

$$P = \{p_1, p_2, \ldots, p_j\}$$

$$V = \{v_1, v_2, \ldots, v_k\}$$

$$|E| = |A| * |P| * |V|$$

Next, we consider a set of possible descriptions (*D*) that could be used for some target event *e*. Each of these descriptions might also apply to other events in the possible set; the number of events some description *d* can refer to is notated as |*d*|. We assume that there is a single, unambiguous label available for each agent, patient, and verb. This means that for a single-word description, |*d*| can be defined easily: for instance, the word referring to the agent can refer to any of the events in E that include that agent plus some patient and verb:

[3] We tested just two values for the number of verbs (*k*): 5 and 50; we use *k*=5 in all models. The effect of increasing *k* is to increase the relative likelihood of including the verb in an utterance.

(4)

$$d = \text{"}A\text{"} \rightarrow |d| = j * k$$

For these models, we consider the set of utterances *D* = ["AV", "VP", "AP"] (i.e. two words, produced in any order). A two-word description like "AV" can refer to any event in E that contains that agent, that verb, and then some patient:

(5)

$$d = \text{"}AV\text{"} \rightarrow |d| = k$$

All of the following models use these same assumptions about the relationship between a particular context (containing possible agents, patients, and verbs) and the number of events a particular utterance could refer to; they differ in how utterances are produced given this information. The outputs of the models are shown in Figure 3.

'Non-Pragmatic' (Cost only) model (p_{NP}): We begin with a baseline model that does not take any aspect of the context in which an utterance was produced into account. In our dataset, we found an overall difference in the frequency with which certain utterances are produced (in particular, AV sequences are less frequent than VP sequences), averaging across contexts. We can consider this as reflecting a differential cost to the speaker of producing each of these utterances. The corresponding model just produces utterances at this baserate (p_{human}), with no effects of the context in which the utterance is produced. Because the human participants sometimes produced a word that did not refer to the agent, verb, or patient, we estimate these probabilities for the model by re-normalizing over only the 'standard' responses (~84% of all

reponses). The probability of producing each two-word description *d* for this model is simply this global likelihood:

(6)

$$p_{CM}(d|e,E) = p_{CM}(d) = \begin{Bmatrix} p_{human}(AV) \text{ if } d = "AV" \\ p_{human}(VP) \text{ if } d = "VP" \\ p_{human}(AP) \text{ if } d = "AP" \end{Bmatrix}$$

To avoid overfitting, we evaluate this model by randomly splitting the human data in half to calculate these parameters from the data, and evaluating each set of predictions against the *other* half of the human dataset.

Pragmatic 'succeed/fail' heuristic (p_{SF}): Many common-ground type experiments (e.g. Sedivy 1999) tacitly assume that an utterance is pragmatically helpful when it uniquely identifies the target referent. For this model we assume the verb is always mentioned because (unlike the possible agents and patients), there is no direct information about the verb present in the context image, and therefore this word will always be highly informative. We thus assume that the possible utterances are simply (unordered) **AV** or **VP,** with the probabilities of producing the two utterances summing to one. The probability of each utterance is given by:

(7)

$$p_{SF}(d|e,E) = \frac{\begin{Bmatrix} 1 \; if \; |d| = 1 \\ \varepsilon \; if \; |d| > 1 \end{Bmatrix}}{\sum_{d' \in D} p(d'|e,E)}$$

That is, out of the available descriptions, the model will consider only the ones that succeed in uniquely identifying the target event. The ε symbol represents a small number arbitrarily close to zero, and indicates that this model is extremely unlikely to choose the

uninformative utterance if any informative ones are available. As shown in equations (8-9), its exact value does not impact the quantitative predictions the model makes. If there is a single informative choice (|*d*|=1), the model will select it approximately deterministically:

(8)

$$p_{SF}(d_{informative}|\, e, E) = \frac{1}{1+\varepsilon} \cong 1$$

But if neither utterance is fully informative (that is, if |d| > 1), the two utterances are produced at chance[4]:

(9)

$$p_{SF}(d_{uninformative}|e, E) = \frac{\varepsilon}{\varepsilon+\varepsilon} \cong \frac{1}{2}$$

Rational speaker (p_{RS}): We implement Frank & Goodman's RSA model, which states that a description *d* will be chosen in inverse proportion to how many events that description can apply to[5]. Thus the probability for each of the three possible 2-word descriptions is:

(10)

$$p_{RS}(d|e, E) = \frac{|d|^{-1}}{\sum_{d' \in D}|d'|^{-1}}$$

[4] This is also the case if both utterances are informative, though in this experiment it is never the case that *both* **AV** and **VP** would uniquely identify the event.

[5] The RSA model also includes a prior probability term (i.e. how often each event is expected to occur); here we assume that all the events have equal prior probability of occurring.

Model comparison: To facilitate comparison with the human results (Figure 3) we plot the probabilities that a word for a particular element (A, V, P) *is included* in the utterances generated by each model. The 'non-pragmatic' model that considers only base rate performs relatively poorly at matching human performance ($r(36) = 0.63$, this and all model comparison p values are < 0.0001; we randomly divided the human data into two halves to avoid overfitting to parameters estimated from the data). The Succeed/Fail model somewhat better ($r(36) = 0.75$), and the rational speaker model better still ($r(36) = 0.81$). In the supplementary section, we compare versions of the latter two models that also incorporate information about the baserate of each words; again, the rational-speaker model is a closer fit to human performance than the equivalent succeed/fail model.

Discussion

As predicted by the RSA, when participants described events after seeing arrays of possible agents and patients, their two-word answers reflected the degree to which a given word could convey new information about the event. Participants were more likely to mention the agent of the event when the agent was more ambiguous, and more likely to mention the patient when the patient was more ambiguous. This was not limited to cases where an event could be uniquely identified: even for the intermediate cases where there were multiple agents *and* multiple patients in the array, participants still chose the two-word sequence that reduced uncertainty the most. Quantitative comparison to the RSA reveals a close fit to human data, with a baseline-adjusted version of the RSA performing best.

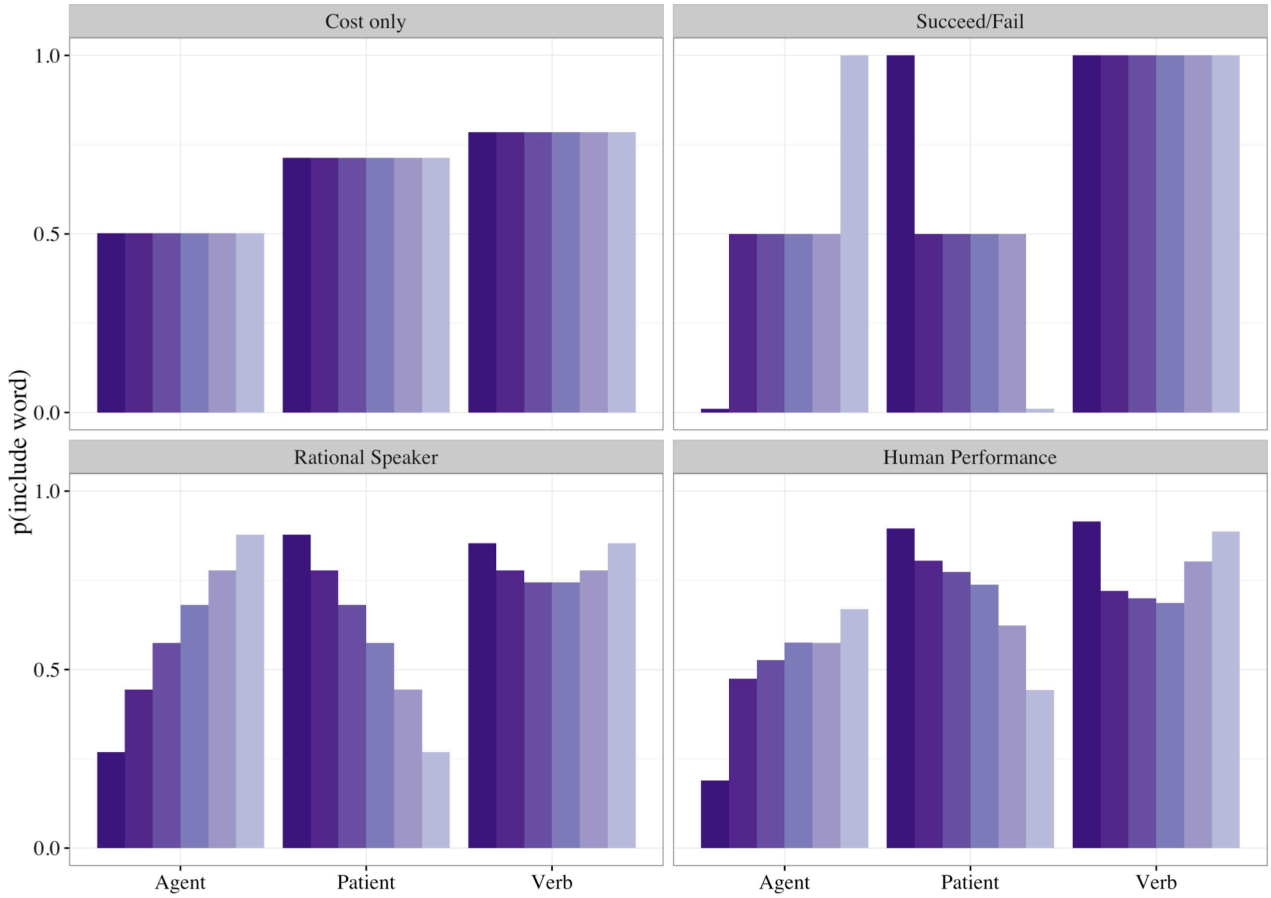


Figure 3: Predictions of the generative models for including Agent, Patient, and Verb in event descriptions; human performance (reproduced from Figure 2) is shown at the bottom right for comparison.

While understanding language appears to involve assuming that we are listening to rational speakers, our own speech also involves messy, sometime under- or over-informative utterances. Nevertheless, we mainly succeed in getting our meanings across, and it is clear that at least some aspects of adult speech are well designed for robustly transferring information. While there is a rich literature on how speakers accommodate non-linguistic context when describing individual objects (c.f. Brennan & Clark, 1996; Brown-Schmidt & Tanenhaus, 2008; Nadig & Sedivy, 2002; Pogue, Kurumada, & Tanenhaus, 2016; Sedivy, 1999), this study provides the first evidence that adults' language production is affected, at the level of argument structure, by the

graded informativity of possible utterances in context. Although the two kinds of shortened sentences (Agent-Verb, e.g., GIRL READ, and Verb-Patient, e.g., READ BOOK) are on average equal in length and express the same amount of information, participants recognize that informativity depends on the set of possible alternative events. This holds even when either utterance will leave some ambiguity, suggesting that RSA-type listeners are correct: their speaker partners are choosing what to say and what to omit in a way that can maximally reduce their uncertainty.

Understanding how listeners and speakers represent contexts and possible messages for verbs and events is a puzzling problem. In noun-referent studies, participants (listeners or speakers) need simply note how many possible referents there are and what features differ between them (e.g. Stiller, Goodman, & Frank, 2015). For sentence level meanings, the set of possible messages is much larger than the number of visible referents. When there are three potential agents and four patients, there are twelve possible combinations, and there may often be multiple verbs under consideration. Beyond this, the listener might have to guess at likely events, as well as multiple ways of referring to that event: beyond just relations between a *girl* and an *apple*, speakers and listeners must consider the many different propositions or perspectives that can be used to refer to the same event: (e.g., a girl *swinging a bat* and *hitting a ball towards the outfielder* can describe the very same event; cf. Gleitman, 1990; Kline, Snedeker, & Schulz, 2016). These perspectives might differ in argument structure, so that a listener might need to consider multiple argument sets: an agent and patient, an agent, theme, and recipient, and so on. Furthermore, in the real world many referents, especially humans, can play many roles (e.g., agent and patient of *hugging*), and some possible referent pairs will permit different interactions due to either selectional restrictions or real-world knowledge. We may be able to use the current

paradigm to address features of argument structure communication like these: if a speaker learns that *wugging* can be performed by animals but not people, will they take this information into account when designing utterances for a partner who does or doesn't know this restriction? How far do parallels between messages about object identity and propositions about the world (event descriptions) extend? Which of the complexities of sentence-level predictability do speakers and listeners fold into their models of communicative context? Understanding the dynamics of utterance production in these contexts will further our understanding of how adults calculate and use informativity to accomplish our communicative goals.

Acknowledgements

We would like to thank the members of the Schulz and Gibson labs for their helpful feedback; Audra Podany, Olivia Murton, and Dmetri Hayes for assistance in creating stimuli and data annotation, and all of the participating AMT workers for their involvement in the study. This work was funded by grants from the National Science Foundation to Edward Gibson and Melissa Kline.

References

Altmann, G. T. ., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *71*, 247–263.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Blyth, C. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338).

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Brennan, S. E., & Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, *1*(2), 274–291. https://doi.org/10.1111/j.1756-8765.2009.01019.x

Brooks, J., Groening, M., Jean, A., Scully, M., Sakal, R., & 20th Century Fox Home Entertainment. (2007). *The Simpsons movie (Website)*. Retrieved from http://web.archive.org/web/20120317090828/http://www.simpsonsmovie.com

Brown-Schmidt, S., & Tanenhaus, M. (2008). Real-Time Investigation of Referential Domains in Unscripted Conversation: A Targeted Language Game Approach. *Cognitive Science: A Multidisciplinary Journal*, *32*(4), 643–684. https://doi.org/10.1080/03640210802066816

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51. https://doi.org/10.1016/j.jml.2009.09.002

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. *Logic, Language and Computation*, *1*.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Goodman, N., & Frank, M. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Grice, H. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58.

Horton, W. S. (2005). Conversational Common Ground and Memory Processes in Language Production. *Discourse Processes*, *40*(1), 1–35. https://doi.org/10.1207/s15326950dp4001_1

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Kao, J. T., Levy, R., & Goodman, N. D. (2013). The funny thing about incongruity: A computational model of humor in puns.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007. https://doi.org/10.1073/pnas.1407479111

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The Egocentric Basis of Language Use: Insights From a Processing Approach. *Current Directions in Psychological Science*, *7*(2), 46–50. https://doi.org/10.1111/1467-8721.ep13175613

Kline, M., Snedeker, J., & Schulz, L. (2016). Linking Language and Events: Spatiotemporal Cues Drive Children's Expectations About the Meanings of Novel Transitive Verbs. *Language Learning and Development*, 1–23. https://doi.org/10.1080/15475441.2016.1171771

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178. https://doi.org/10.1016/j.jml.2015.03.003

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*(3), 550–557. https://doi.org/10.3758/BF03196312

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00226

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/S0140525X04000056

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-Specific Generalization of Pragmatic Inferences based on Under- and Over-Informative Prenominal Adjective Use. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02035

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Sedivy, J. (1999). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*.

Shannon, C. E. (1949). Communication in the Presence of Noise. *Proceedings of the IRE*, *37*(1), 10–21. https://doi.org/10.1109/JRPROC.1949.232969

Stiller, A. J., Goodman, N., & Frank, M. C. (2015). Ad-hoc Implicature in Preschool Children. *Language Learning and Development*, *11*(2), 176–190. https://doi.org/10.1080/15475441.2014.927328

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

van Son, R. J. J. H., & van Santen, J. P. H. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1–2), 100–123. https://doi.org/10.1016/j.specom.2005.06.005

Weel, J. (2008). *Willow, a Python library for economics*. Retrieved from http://econwillow.sourceforge.net/