Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions

Caroline Graf, Judith Degen, Robert X.D. Hawkins, Noah D. Goodman
cgraf@uos.de, {jdegen,rxdh,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Nominal reference is very flexible—the same object may be called *a dalmatian*, *a dog*, or *an animal* when all are literally true. What accounts for the choices that speakers make in how they refer to objects? The addition of modifiers (e.g. *big dog*) has been extensively explored in the literature, but fewer studies have explored the choice of noun, including its level of abstraction. We collected freely produced referring expressions in a multi-player reference game experiment, where we manipulated the object's context. We find that utterance choice is affected by the contextual informativeness of a description, its length and frequency, and the typicality of the object for that description. Finally, we show how these factors naturally enter into a formal model of production within the Rational Speech-Acts framework, and that the resulting model predicts our quantitative production data. **Keywords:** referential expressions, levels of reference, basic level, experimental pragmatics, computational pragmatics

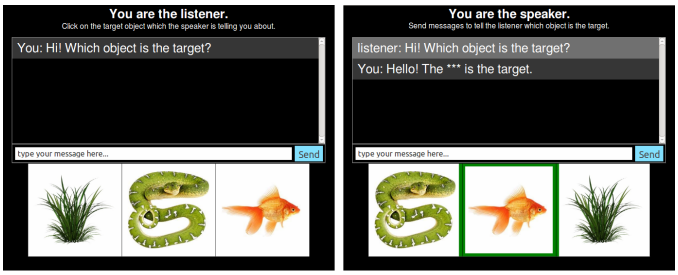

Figure 1: Screenshots from speakers' and listeners' points of view, showing role names and short task descriptions, the chatbox used for communication and a display of three pictures of objects. The referent was identified to the speaker by a green box.

Referring to objects is a core function of human language, and a wealth of research has explored how speakers choose referring expressions (Herrmann & Deutsch, 1976; Pechmann, 1989; van Deemter, Gatt, van Gompel, & Krahmer, 2012). However, most of this literature has focused on the addition of modifiers (as in the choice between "the dog", "the brown dog", and "the big brown dog", e.g., Sedivy, 2003; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Here we investigate how speakers choose a simple nominal referring expression—what governs the choice of calling a particular object "the dalmatian", "the dog", or "the animal" when all are literally true? That is, what governs the choice of the taxonomic level at which an object is referred to? Noun choice can be seen as the most basic decision in forming a referring expression. Like modification, these choices differ in their specificity; unlike modification, the number of words used does not differ—in English, *some* noun must be chosen. In this paper we provide experimental evidence from a coordination game regarding the flexible choice of nominal referring expressions and explain this data with a probabilistic model of pragmatic production.

Previous evidence about the generation of referring expressions suggests that choice of reference level will depend on the interplay of several factors. Grice's Maxim of Quantity (Grice, 1975) implies a pressure for speakers to be sufficiently *informative*. For instance, a speaker who is trying to distinguish a dalmatian from a German Shepherd would be expected to avoid the insufficiently specific term "dog" (Brennan & Clark, 1996). On the other hand, recent work in experimental pragmatics has shown that the choice of referring expression depends on the *cost* of utterance alternatives (Rohde, Seyfarth, Clark, Jäger, & Kaufmann, 2012; Degen, Franke, & Jäger, 2013); sometimes, speakers are willing to produce a cheap ambiguous utterance rather than a costly (e.g. long or difficult-to-retrieve) unambiguous one. Finally, classic work on concepts suggests that *typicality* of a referent within its category affects the choice of reference (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). In particular, speakers will generally choose to refer at the *basic level* (e.g. "dog"), but may become more specific for objects that are atypical for the basic level term.

To evaluate the impact of these factors on nominal reference we constructed a two-player online game (Fig. 1). Participants saw a shared context of objects, one of which was indicated as the referent only to the speaker. The speaker was asked to communicate this object to the listener, who then chose among the objects. Critically, the speaker and listener communicated by free use of a chat window, allowing us to gather relatively natural referring expressions. We manipulated the category of distractor objects and used items that varied in utterance complexity and object typicality. This allowed us to evaluate whether each factor influences the referring expressions generated by participants. We expect that speakers will (1) tend to avoid longer or less frequent terms, and (2) will pragmatically prefer more specific referring expressions when the target and distractor(s) belong to the same higher-level taxonomic category or when distractors are more typical members of that category level.

A promising modeling approach for capturing the quantitative details of human language use is the Rational Speech-Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). The RSA framework has been applied to many language interpretation tasks (e.g. Goodman & Stuhlmüller, 2013; Kao, Wu, Bergen, & Goodman, 2014), but relatively rarely to production data (but see Franke,

2014; Orita, Vornov, Feldman, & Daumé III, 2015). We describe an RSA model of nominal reference that includes informativeness, cost, and typicality effects. A speaker in RSA is treated as an approximately optimal decision maker who chooses which utterance to use to communicate to a listener. The speaker has a utility which includes terms for the cost of producing an utterance (in terms of length or frequency) and the informativeness of the utterance for a listener. The listener is treated as a literal Bayesian interpreter who updates her beliefs given the truth of the utterance. These truth values are usually treated as deterministic (an object either is a "dog" or it is not); here we relax this formulation in order to incorporate typicality effects. That is, we elicit typicality ratings in a separate experiment, and model the listener as updating her beliefs by weighting the possible referents according to how typical each is for the description used. We evaluate the quantitative model predictions against our production data. The model also allows us to evaluate the need for each extra component—typicality, length, frequency—and determine whether the empirical bias toward reference at the basic level (Rosch et al., 1976) can be accounted for without building it in as a separate factor.

Experiment: nominal reference game

Methods

Participants and materials We recruited 56 self-reported native speakers of English over Mechanical Turk. Participants completed the experiment in pairs of two, yielding 28 speaker-listener pairs.

Stimuli were selected from nine distinct domains, each corresponding to distinct basic level categories such as "dog." For each domain, we selected four subcategories to form our target set (e.g. "dalmatian", "pug", "German Shepherd" and "husky"). Each domain also contained an additional item which belonged to the same basic level category as the target (e.g. "greyhound") and items which belonged to the same supercategory but not the same basic level (e.g. "elephant" or "squirrel"). The latter items were used as distractors.

Each trial consisted of a display of three images, one of which was designated as the target object. Every pair of participants saw every target exactly once, for a total of 36 trials per pair. These target items were randomly assigned distractor items which were selected from four different context conditions, corresponding to different communicative pressures (see Fig. 2). We refer to these conditions with pairs of numerals specifying which levels of the taxonomy are present in the distractors: (a) **item12**: one distractor of the same basic level and one distractor of the same superlevel (e.g. target: "dalmatian", distractor 1: "greyhound", distractor 2: "squirrel"), (b) **item22**: two distractors of the same superlevel, (c) **item23**: one distractor of the same superlevel and one unrelated item and (d) **item33**: two unrelated items.

Furthermore, the experiment contained 36 filler items, in which participants were asked to produce referential expressions for objects which differed only in size and color. Images

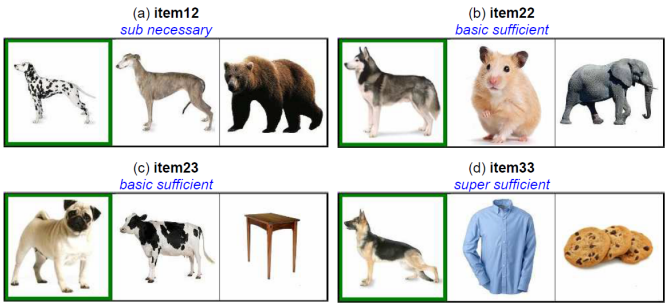

Figure 2: The four context conditions, exemplified by the *dog* domain. The target is outlined in green; the types of distractors differ with condition (see text).

from filler trials were not reused on target trials. Trial order was randomized.

Procedure Pairs of participants were connected through a real-time multi-player interface (Hawkins, 2015), with one member of each pair assigned the speaker role and the other to the listener role. Participants kept their allotted roles for the entire experiment. The setup for both the speaker and the listener is shown in Fig. 1. Each saw the same set of three images, but positions were randomized to rule out trivial position-based references like "the middle one." The target object was identified by a green square surrounding it for the speaker (but not listener). Players used a chatbox to send text messages to each other. The task was for the speaker to get the listener to select the target object.

Annotation To determine the level of reference for each trial, we followed the following procedure. First, trials on which the listener selected the wrong referent were excluded, leading to the elimination of 1.2% of trials. Then, speakers' and listeners' messages were parsed automatically; the referential expression used by the speaker was extracted for each trial and checked for whether it contained the current target's correct sub, basic or super level term using a simple grep search. In this way, 66.2% of trials were labelled as mentioning a pre-coded level of reference. In the next step, remaining utterances were checked manually to determine whether they contained a correct level of reference term which was not detected by the parsing algorithm due to typos or grammatical modification of the expression. In this way, meaning-equivalent alternatives such as "doggie" for "dog", or contractions such as "gummi","gummies" and "bears" for "gummy bears" were counted as containing a level of reference term. This caught another 13.8% of trials. A total of 20.0% of correct trials were excluded because the utterance consisted only of an *attribute* of the superclass ("the living thing" for "animal"), of the basic level ("can fly" for "bird"), of the subcategory ("barks" for "dog") or of the particular instance ("the thing facing left") rather than a category noun. These kinds of attributes were also sometimes

mentioned in addition to the noun in the trials which were included in the analysis—4.0% of sub level terms, 12.6% of basic level terms, and 46.2% of super level terms contained an additional modifier. On 0.5% of trials two different levels of reference were mentioned; in this case the more specific level of reference was counted as being mentioned in this trial.

Typicality norms To examine the influence of typicality on speaker behavior, we obtained typicality estimates in a separate norming study. 240 participants were recruited through Mechanical Turk. On each trial, we presented participants with an image from the main experiment and asked them "How typical is this for X?", where X was a category label at the sub-, basic-, or super- level. They then adjusted a slider bar ranging from *not at all typical* to *very typical*.

Due to the large number of possible combinations of objects, we only collected norms for certain combinations of objects and descriptions: for each target (e.g., dalmatian), we collected typicality at all three levels ("dalmatian," "dog," and "animal"). For each distractor of the same superclass as the target (*distsamesuper*, e.g., a kitten), we collected typicality at all three levels of the *target*. For each distractor of a different superclass (*distdiffsuper*, e.g., a basketball) we only collected typicality at the super- level of the target ("animal") and assumed lowest typicality at the other levels. This resulted in the following distribution of 745 norms: *target-sub* (36), *target-basic* (36), *target-super* (36), *distdiffsuper-super* (168), *distsamesuper-sub* (331), *distsamesuper-basic* (93), and *distsamesuper-super* (45).

Each participant provided typicality ratings for 7 *target*, 10 *distdiffsuper*, and 28 *distsamesuper* cases (randomly sampled from the total set of items). Each case received between 6 and 27 ratings. Raw slider values ranged from 0 (not typical) to 1 (very typical); average slider values were used as the typicality values throughout our results.

Results

Proportions of sub, basic, and super level utterance choices in the different context conditions are shown in the top row of Fig. 3. The sub level term was preferred where it was necessary for unambiguous referent identification, i.e., when a distractor of the same basic level category as the target was present in the scene (item12, e.g. target: dalmatian, distractor: greyhound). Where it was not necessary (i.e., when there was no other object of the same basic level category present, as in conditions item22, item23 and item33), there was a clear preference for the basic level term. The super level term was strongly dispreferred overall, though it was used on some trials, especially where informativeness constraints on utterance choice were weakest (item33).

To test for the independent effects of informativeness, length, frequency, and typicality on sub-level mention, we conducted a mixed effects logistic regression. Frequency was coded as the difference between the sub and the basic level's log frequency, as extracted from the Google Books Ngram

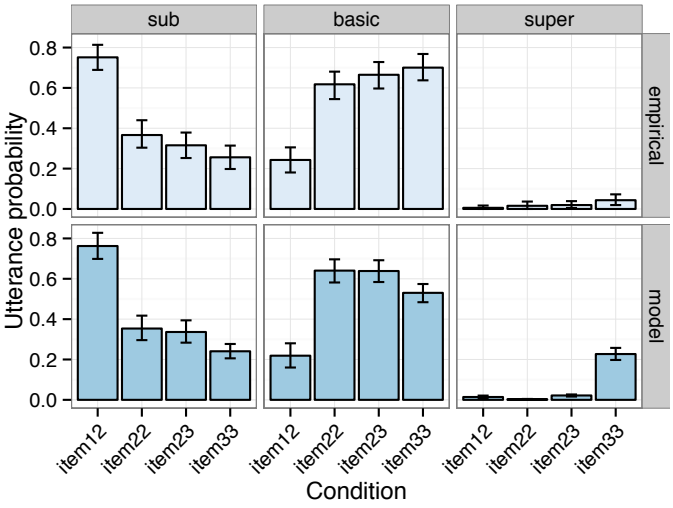

Figure 3: Empirical utterance probabilities (top row) and model posterior predictive MAP estimates (bottom row) by condition, collapsed across targets and domains. Error bars indicate bootstrapped 95% confidence intervals.

English corpus ranging from 1960 to 2008. Length was coded as the ratio of the sub to the basic level's length.[1] That is, a higher frequency difference indicates a *lower* cost for the sub level term compared to the basic level, while a higher length ratio reflects a *higher* cost for the sub level term compared to the basic level.[2] Typicality was coded as the ratio of the target's sub to basic level label typicality. That is, the higher the ratio, the more typical the object was for the sub level label compared to the basic level. For instance, the panda was relatively atypical for its basic level "bear" (mean rating 0.75) compared to the sub level term "panda bear" (mean rating 0.98), which resulted in a relatively *high* typicality ratio.

Condition was coded as a three-level factor: *sub necessary*, *basic sufficient*, and *super sufficient*, where item22 and item23 were collapsed into *basic sufficient*. Condition was Helmert-coded: two contrasts over the three condition levels were included in the model, comparing each level against the mean of the remaining levels (in order: *sub necessary*, *basic sufficient*, *super sufficient*). This allowed us to determine whether the probability of type mention for neighboring conditions were significantly different from each other, as suggested by Fig. 3.[3] The model included random by-speaker and by-domain intercepts.

A summary of results is shown in Table 1. The log odds

[1] We used the mean empirical lengths in characters of the utterances participants produced. For example, the minivan, when referred to at the subcategory level, was sometimes called "minivan" and sometimes "van" leading to a mean empirical length of 5.64. This is the value that was used, rather than 7, the length of "minivan".

[2] We replicate the well-documented negative correlation between length and log frequency ($r = -.53$ in our dataset).

[3] Adding terms that code the ratio of the sub vs super level frequency and length did not lead to an improvement of model fit.

Table 1: Mixed effects model summary.

	Coef β	SE(β)	p
Intercept	−0.30	0.35	>0.4
Condition sub.vs.rest	2.46	0.24	<**.0001**
Condition basic.vs.super	0.52	0.23	<**.05**
Length	−0.52	0.14	<**.001**
Frequency	−0.02	0.08	>0.78
Typicality	4.17	0.84	<**.0001**
Length:Frequency	−0.30	0.11	<**.01**

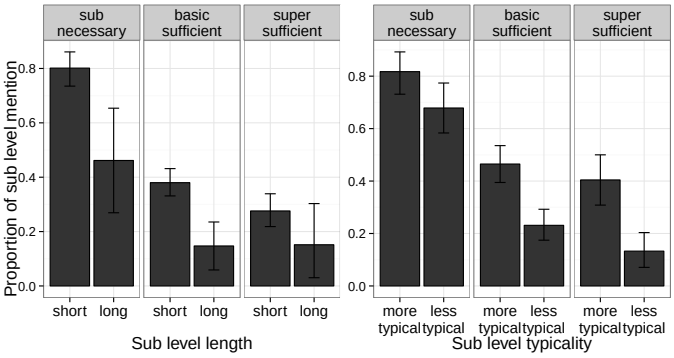

Figure 4: Probability of using sub, basic and super level terms. Left: when the sub length is relatively short (.67,2] or long [2,4.67) compared to the basic level term length. Right: when the target object was relatively more [1.06,1.91) or less (.88,1.06] typical for the sub compared to the basic level term.

of mentioning the sub level term was greater in the *sub necessary* condition than in either of the other two conditions, and greater in the *basic sufficient* condition than in the *super sufficient* condition, suggesting that the contextual informativeness of the sub level mention has a gradient effect on utterance choice.[4] There was also a main effect of typicality, such that the sub level term was preferred for objects that were more typical for the sub level compared to the basic level description (Fig. 4). In addition, there was a main effect of length, such that as the length of the sub level term increased compared to the basic level term ("chihuahua"/"dog" vs. "pug"/"dog"), the sub level term was dispreferred ("chihuahua" is dispreferred compared to "pug", Fig. 4). Finally, while there was no main effect of frequency, we observed a significant length by frequency interaction, such that there was a frequency effect for the relatively shorter but not the relatively longer sub level cases: for shorter sub level terms, relatively high-frequency sub level terms were more likely to be used than relatively low-frequency sub level terms.

Unsurprisingly, there was also significant by-participant and by-domain variation in the log odds of sub level term mention. For instance, mentioning the subclass over the basic level term was preferred more in some domains (e.g. in the "candy" domain) than in others. Likewise, some domains had a greater preference for basic level terms (e.g. the "shirt" domain). Using the superclass term also ranged from hardly being observable (e.g. the "flower" domain) to being used more frequently (e.g. in the "bird" domain). Nevertheless, mentioning the sub level term was always the most frequent choice where a distractor of the same basic level was displayed. Furthermore, it was the case in all domains that the sub level term was mentioned most frequently and the basic level least frequently in just this condition, compared to the other three conditions.

Modeling level of reference

We formulated a probabilistic model of reference level selection that integrates contextual informativeness, utterance cost, and typicality. As in earlier Rational Speech-Acts (RSA) models (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), the speaker seeks to be informative with respect to an internal model of a literal listener. This listener updates her beliefs to rule out possible worlds that are inconsistent with the meaning of the speaker's utterance. Rather than assuming that words have deterministic truth conditions, as has usually been done in the past, we account for typicality by allowing each label a graded meaning. For instance, the word "dog" describes a dalmatian better than a grizzly bear, but it also describes a grizzly bear better than a tennis ball. The speaker also seeks to be parsimonious: the speaker utility includes both informativeness and word cost; cost includes both length and frequency.

Formally, we start by specifying a literal listener L_0 who hears a word l at a particular level of reference in the context of some set of objects O and forms a distribution over the referenced object, $o \in O$:

$$P_{L_0}(o|l) \propto [\![l]\!](o).$$

Here $[\![l]\!](o)$ is the lexical meaning of the word l when applied to object o. We take this to be a real number indicating the degree of acceptability of object o for category l. We relate this to our empirically elicited typicality norms via an exponential relationship: $[\![l]\!](o) = \exp(\text{typicality}(o,l))$.[5] This relationship is motivated by considering the effect of a small difference in typicality on choice probability: in our elicitation experiment a small difference in rating should mean the same thing at the top and bottom of the scale (it is visually equivalent on the slider that participants used). In order for a small difference in typicality rating to have a constant effect on relative choice probability (which is a ratio), the relationship must be exponential.

Next, we specify a speaker S_1 who intends to refer to a particular object $o \in O$ and chooses among possible nouns $l \in$

[4] Importantly, model comparison between the reported model and one that subsumes basic and super under the same factor level revealed that the three-level condition variable is justified ($\chi^2(1) = 5.7$, $p < .05$), suggesting that participants don't simply revert to the basic level unless contextually forced not to.

[5] Cases where typicality was not elicited were assumed to have typicality 0.

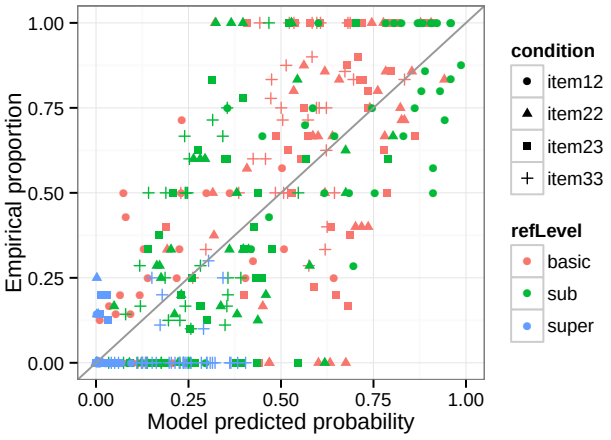

Figure 5: Mean empirical production data for each level of reference against the MAP of the model posterior predictive at the by-target level.

$\mathcal{L}(o)$. We take $\mathcal{L}(o)$ to be the three labels for o at sub, basic, and super level. The speaker chooses among these nouns in a way that is influenced by informativeness of the noun for the literal listener ($\ln P_{L_0}(o|l)$), the frequency ($\hat{c}_f$) and the length ($\hat{c}_l$), each weighted by a free parameter:

$$P_{S_1}(l|o) \propto \exp(\lambda \ln P_{L_0}(o|l) + \beta_f \hat{c}_f + \beta_l \hat{c}_l)$$

Length cost $\hat{c}_l$ was defined as the empirical mean number of characters used to refer at that level and frequency cost $\hat{c}_f$ was the log frequency in the Google Books corpus from 1960 to the present.

We performed Bayesian data analysis to generate model predictions, conditioning on the observed production data (coded into sub, basic, and super labels as described above) and integrating over the three free parameters. We assumed uniform priors for each parameter: $\lambda \sim Unif(0,20)$, $\beta_f \sim Unif(0,5)$, $\beta_l \sim Unif(0,5)$. We implemented both the cognitive and data-analysis models in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, electronic). Inference for the cognitive model was exact, while we used Markov Chain Monte Carlo (MCMC) to infer posteriors for the three free parameters.

Point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives at the target level (collapsing across distractors for each target, within each condition) are compared to empirical data in Fig. 5. On the by-target level the model achieves a correlation of $r = .79$. Looking at results on the by-domain level (collapsing across targets) and on the by-condition level (further collapsing across domains, as in Fig. 3) yields correlations of .88 and .96, respectively. The model does a good job of capturing the quantitative patterns in the data, especially considering the sparsity of our data at the by-target level. One clear flaw is that the model predicts greater use of the super level label than people exhibit. Further systematic deviation appears likely for specific

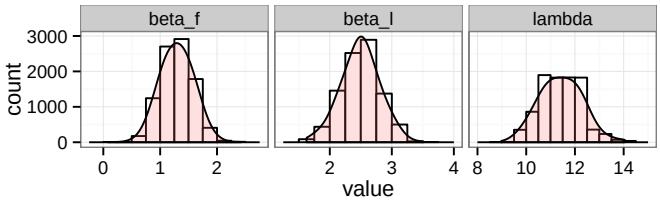

Figure 6: Posterior distribution over model parameters. Maximum a posteriori (MAP) $\lambda = 10.8$, 95% highest density interval (HDI) $= [9.7, 12.8]$; MAP $\beta_l = 2.5$, HDI $= [1.9, 3.1]$; MAP $\beta_f = 1.3$, HDI $= [0.8, 1.8]$.

items. On examination, candy items like "gummy bears" or "jelly beans" were particularly problematic, being referred to primarily by their sub level term in all contexts.

Parameter posteriors are presented in Fig. 6. Informativeness is weighted relatively strongly, while length is weighted somewhat more strongly than frequency. Note that the 95% highest density intervals (HDIs) for all three weight parameters exclude zero, indicating that some contribution of each is useful in explaining the data. In order to ascertain whether typicality was indeed contributing to the explanatory power of the model, we ran an additional Bayesian data analysis with an added typicality weight parameter $\beta_t \in [0,1]$. This parameter interpolated between empirical typicality values (when β_t=1) and deterministic (i.e. 0 or 1) *a priori* values based on the true taxonomy (when β_t=0). We found a MAP estimate for β_t of .94, HDI $= [0.88, 1]$, strongly indicating that it is useful to incorporate empirical typicality values. Finally, we ran a model including a parameter weighting the *product* of frequency and cost, corresponding to the interaction term in our regression analysis. Its posterior distribution was strongly peaked at 0, indicating that any contribution of the interaction is already captured by other aspects of the model.

Discussion and conclusion

The choice speakers make of how to refer to an object is influenced by a rich variety of factors. In this paper, we specifically investigated the choice of level of reference in nominal referring expressions. In an interactive reference game task in which speakers freely produced referring expressions, utterance choice was affected by utterance cost (in terms of length and frequency), contextual informativeness (as manipulated via distractor objects), and object typicality. The interplay of these factors is naturally modeled within the RSA framework, where speakers are treated as choosing utterances by soft-maximizing utterance utility, which includes terms for informativeness and cost. In previous formulations of RSA models, informativeness was determined by a deterministic semantics; here we "softened" the semantics by allowing nouns to apply to objects to the extent that those objects were rated as typical for the nouns. The resulting model provided a good fit to speakers' empirical utterance choices, both qualitatively and quantitatively.

The model predicts a well-documented preference for

speakers to refer to objects at the basic level when not constrained by contextual considerations (Rosch et al., 1976). In our model, this preference emerges naturally from cost considerations: basic-level labels tend to be shorter and more frequent than sub and super level terms. However, speakers did not always use the basic level term, even when unconstrained by context. In certain cases where object typicality was relatively high for the sub level term compared to the basic level term, that term was preferred (as was the case for "panda bear"), suggesting an interesting interplay between typicality and level of description. While our results show that a model can capture several basic-level phenomena through frequency, length, and typicality features, it leaves open the origin and causal role of these linguistic regularities. Future research will be needed to determine how linguistic regularities are related to conceptual regularities and why.

An interesting analogy can be drawn from choosing a noun to choosing a set of adjectives; that is, between selection of a level of reference in simple nominal referring expressions and selection of a set of features to include in modified referring expressions. For the latter, a much discussed phenomenon is that of *overinformative* modifier use (Gatt, Krahmer, van Deemter, & van Gompel, 2014)—for example, saying "big blue" when all objects in the context are blue. The preference for the basic level in the *super sufficient* condition and the still substantial use of sub level terms in the *basic sufficient* condition can also be considered overinformative. However, we showed that a Rational Speech-Acts model using non-deterministic semantics, derived from typicality estimates, predicts that speakers *should* use these more specific descriptions. The extent to which similar considerations may apply to modified referring expressions should be explored. Future research should also examine the interaction of these choices: circumstances under which speakers choose a modifier and how nominal and modifier choice interact.

Acknowledgments

This work was supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and an SNF Early Postdoc. Mobility Award to JD. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Brennan, S. E., & Clark, H. H. (1996, nov). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, *22*(6), 1482 – 1493.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. In *Proceedings of the 35th annual conference of the cognitive science society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society.*

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, *29*(8), 899–911.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

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

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Huber.

Kao, J., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12002–12007.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Orita, N., Vornov, E., Feldman, N., & Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1639–1649.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. In *Proceedings of the 16th workshop on the semantics and pragmatics of dialogue* (pp. 107 – 116).

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Sedivy, J. C. (2003, jan). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3–23.

van Deemter, K., Gatt, A., van Gompel, R. P. G., & Krahmer, E. (2012, apr). Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–83.