The special status of color in pragmatic reasoning: evidence from a language game

Peter Baumann (baumann@u.northwestern.edu)
Department of Linguistics
Northwestern University, Evanston, IL, USA

Abstract

In current approaches to pragmatic reasoning the comprehension and production of referring expressions is modeled as a result of the interlocutors' mutual perspective-taking. While such models of pragmatic reasoning have been empirically validated in referential language games experiments, empirical (and computational) work on the generation of referring expressions has shown that speakers do not always take the listener's perspective into account, but instead produce referring expressions according to their own preferences. One particularly well studied example is color: speakers often include color terms in their referring expressions even if they do not help identify the intended referent. We show that like speakers, listeners treat color differently from other properties like e.g. size. Our results suggest that listeners do not seem to perform much pragmatic reasoning when the referring expression only expresses color, but instead follow a simple salience-based heuristic.
Keywords: Referring Expressions; Pragmatics; Language games; Language Production; Language Comprehension

Introduction

Reference and referring expressions are central to human communication: in order to refer to an object or person in the world, a speaker needs to produce an appropriate referring expression so that a listener will be able to identify that referent given the expression and the context. A number of attempts have been made to provide a more quantitative understanding of the production and comprehension of referring expressions. In an overly simplifying manner, these attempts may be grouped into models of pragmatic reasoning and models of content selection.

Focusing primarily on comprehension, models of pragmatic reasoning, such as game-theoretic models (e.g. Benz & Van Rooij, 2007; Jäger, 2011) or Bayesian models (e.g. Frank & Goodman, 2012), assume that the speaker and hearer reason about each other's perspectives: the hearer is assumed to interpret a speaker's expression as referring to the referent for which the expression is 'optimal' under the perspective of the speaker, who in turn chooses the referring expression to be 'optimal' under the hearer's perspective, etc. A trivial solution to this recursive reasoning process is for the speaker to choose a referring expression that explicitly mentions *all* features of the intended referent and is thus absolutely unambiguous in the given context. Since such an expression can hardly qualify as efficient, however, the above models make the crucial additional assumption that speakers have a preference for the most economic (i.e. shortest and least effortful) expression.

Focusing primarily on production, models of content selection or the generation of referring expressions (Dale & Reiter, 1995) start from the observation that referring expressions produced by actual speakers are not always 'optimal' and that instead speakers make use of overspecification, i.e. saying more than is strictly necessary (e.g. Pechmann, 1989; Gatt, Krahmer, van Gompel, & van Deemter, 2013; Baumann, Clark, & Kaufmann, 2014). Which properties of a referent are prone to being used in a referring expression even when they do not help to identify it, is assumed to be related to a property's inherent salience: while some properties, most notably a referent's color, are expressed more often than required, other properties, such as size, tend to be used only when necessary to identify a referent (e.g. Gatt, van Gompel, Krahmer, & van Deemter, 2011).

While the inherent salience of referent properties has been shown to influence the production of referring expressions, it remains open if and to what extent it also plays a role in comprehension. In this paper, we provide evidence that the inherent salience of different referent properties, namely color and size, influences the comprehension of referring expressions. More specifically, we show in a referential language game experiment that listeners do not seem to perform much pragmatic reasoning when the referring expression only expresses color, but instead follow a simple salience-based heuristic. When the referring expression expresses size, on the other hand, pragmatic reasoning is more involved and more in line with the predictions of models of pragmatic reasoning.

The remainder of the paper is organized as follows: we first introduce referential language games as an empirical method to study reference. We then review some relevant prior research on pragmatic reasoning and content selection and finally present two experiments, a production and a comprehension experiment whose results are compared to the predictions of a Bayesian model of pragmatic reasoning.

Referential Language Games

Referential language games (Wittgenstein, 1959; Lewis, 1969) have been used to empirically test pragmatic reasoning about referents and referring expressions in (simple) visual contexts. As an example, consider the situation sketched in Figure 1. If in this context, a speaker said *My friend is the one wearing sunglasses*, a listener could infer that she is referring to the person wearing sunglasses and no hat. This follows under the assumption that if the speaker had wanted to refer to the person wearing a hat and sunglasses, she could have used an expression like *My friend is the one wearing a hat*, which unambiguously identifies the intended referent (Grice, 1975). It has been shown that (adult) listeners can easily perform this kind of pragmatic reasoning (e.g. Stiller, Goodman, & Frank, 2011; Degen & Franke, 2012).

In the general form of a language game, speakers and listeners see a visual display of several potential referents, from

Figure 1: Example of the visual context of a language game.

which the speaker picks (or is given) a referent to talk about, while the listener does not know the referent. The speaker then chooses a referring expression, based on which the listener must identify the intended referent.

Language games are particularly suitable for the study of pragmatic reasoning as they allow for an easy manipulation of the depth of (recursive) pragmatic reasoning required for successful communication by changing the distribution of features across the different referents. In Figure 1, the target referent (middle), shares one feature with each of its two competitors: like the competitor to its right, the target is wearing sunglasses, and like the competitor to its left, it is not wearing a hat. So upon hearing a sentence like *the one wearing sunglasses*, the hearer must employ pragmatic reasoning and strengthen the heard utterance to mean *the one wearing sunglasses, but no hat*.

Related Work

In this section, we present one Bayesian model of pragmatic reasoning about referring expressions and then briefly review some relevant studies from the vast literature on the generation of referring expressions.

Models of Pragmatic Reasoning

Models of pragmatic reasoning are based on the assumption that the speaker and hearer reason about each other's perspectives. A Bayesian approach to model a two-level pragmatic reasoning process was proposed by Frank and Goodman (2012): in their model, a listener uses Bayesian inference to infer a speaker's intended referent r given that the speaker used a particular word w:

$$P(r|w) = \frac{P(w|r)P(r)}{\sum_{r}' P(w|r')P(r')} \quad (1)$$

The likelihood $P(w|r)$ of a speaker uttering word w given an intended referent r is assumed to reflect the utility or informativeness of a word used to refer to a particular referent and thus takes the form

$$P(w|r) = \frac{|w|^{-1}}{\sum_{w' \in W} |w'|^{-1}} \quad (2)$$

where $|w|$ denotes the number of referents for which word w is true and W is the set of all words w' that could be used to describe the referent r.

Frank and Goodman (2012) tested this model in a simple referential language game, in which human participants provided probability judgments for the three terms in Equation 1. Participants saw an array of three objects varying along two of the three following property dimensions: shape, color and texture. They were then asked to place bets on

1. which object a speaker is talking about given that she used one of the two words ($P(r|w)$, listener)
2. which object a speaker is talking about given that she used an unknown word ($P(r)$, salience)
3. which (of the two possible) word they would use to refer to a given object ($P(w|r)$, speaker).

Given these estimates, Frank and Goodman (2012) observed a very strong correlation between both the estimated and predicted speaker likelihoods $P(w|r)$ (according to Equation 2) and the estimated and predicted listener probabilities $P(r|w)$ obtained from Equation 1 with $P(w|r)$ according to Equation 2 and the estimated $P(r)$.

Critically, participants were given a probabilistic forced-choice task, as they had to place bets on their answers. However, in the speaker condition, there was no clear option for overspecification, i.e. for using both properties to refer to the given referent. While this design choice of Frank and Goodman (2012) is in line with other experiments on pragmatic reasoning (e.g. Degen & Franke, 2012), it has been shown that when given an a free choice, speakers do make use of overspecification in similar experimental settings (Gatt, van Gompel, van Deemter, & Krahmer, 2013; Baumann et al., 2014).

Overspecification in the Generation of Referring Expressions

The fact that speakers have the option for overspecification and often make use of it has become a well-established fact in the psycholinguistic literature on referring expressions (e.g. Pechmann, 1989; Engelhardt, Bailey, & Ferreira, 2006; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Numerous studies have shown that the properties of referents form preference hierarchies according to their inherent salience and that properties listed high on these hierarchies, such as color, tend to be used by speakers even if they are not necessary to identify an intended referent (Koolen et al., 2011) or if they provide less discriminatory power than a property lower on the hierarchy (Gatt, Krahmer, et al., 2013).

While some of these studies involve rather complex visual contexts from which a specific referent is to be identified, Gatt et al. (2011) report an experiment that closely resembles the design of language games used to test models of pragmatic reasoning: their visual display consisted of three objects, which differed along the two feature dimensions color and size (see Figure 2), and participants were given an open prompt to answer. The results showed that participants made significant use of overspecification, especially if the redundant feature was color. However, the authors of this study

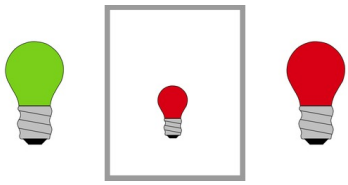

Figure 2: Example of a simple visual context of a language game with size and color. Size is sufficient to identify the target.

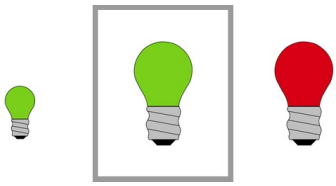

Figure 3: Example of a pragmatic visual context of a language game with size and color.

were primarily interested in the differential effects of color and size on the degree of overspecification, and so the experiment consisted only of conditions in which either color or size or both were unique features of the target object. Crucially, there was no 'pragmatic' condition, in which the target did not have a unique property in context, but a single property could still suffice to identify a referent through pragmatic reasoning.

Experiments

Our experiments add such an 'pragmatic' condition to the one reported by Gatt et al. (2011). More importantly, we also investigate how listeners understand pragmatically 'optimal' utterances in the context of such pragmatic contexts involving properties of different inherent saliences.

As an example of a pragmatic context, consider a visual display like in Figure 3. Here the target referent (middle) shares one property with each of its competitors: it has the same size as the competitor to its right and the same color as the one to its left. Like in the case of the smileys in Figure 1, the target can be identified by referring to only one of its two properties through pragmatic reasoning: if a speaker chooses to refer to the target just by its size (*the large light bulb*), a listener can infer that speaker is more likely to be referring to the large green light bulb than to the small one, since for referring to the latter one she could have used the unique expression *the small light bulb*. By the same line of reasoning the target could also be identified by just referring to its color (*the green light bulb*). So unlike in the smiley case of Figure 1, the array of light bulbs in Figure 3 is symmetrical, i.e. assuming a pragmatic listener, a speaker could refer to the target by either of its two properties.

Experiment 1: Production

Experiment 1 is a production experiment in which we compare speakers referring expressions for referents in'pragmatic' contexts like in Figure 3 with 'simple' contexts, in which the referent has a unique property.

Materials and Design We designed two sets of arrays of three pictures (light bulbs and dinosaurs), like the one in Figure 3. The individual pictures differed in two properties: size and color

The pictures in the arrays were assembled according to three conditions: a *pragmatic condition* (Figure 3) and two *simple conditions*. In the two simple conditions, the target picture (indicated by a grey frame) had a unique property, either *color* or *size* (Figure 2), by which it could easily be referred to (e.g. *small light bulb*).

In the pragmatic condition, the target picture shared one property with each of its two competitors: e.g. in Figure 3, it has the same size as the competitor to its right and the same color as the one to its left. As illustrated above, if speakers were producing pragmatically 'optimal' referring expressions, Figure 3 could be referred to as either *green light bulb* or *large light bulb*.

We employed a two-shot between-subject design: each participant completed only two trials (cf. Frank and Goodman (2012), who also used single trial experiments): one in the pragmatic condition and one in one of the simple conditions. The properties and array sets were counter-balanced across participants and the order of the three pictures was randomized within the array.

Participants and Procedure Using Amazon Mechanical Turk, 96 participants were recruited for the experiment. All participants were naive to the purpose of the experiment and reported to be native speakers of English.

The two trials were presented on a single web page. Each trial consisted of a three-picture array with one picture marked by a grey frame as in Figure 3. Under the picture array there was a text line with the words *Pick the* followed by an open prompt and a period. The participants' task was to fill in the blank so that another person could identify the target picture with the grey frame.

Participants were instructed to imagine that they were communicating with another person over an instant messaging or chat system and that they wanted their partner to pick the picture with the grey frame out of the three pictures in the array. They were told to imagine that their partner saw the same three pictures, but without the frame and in a possibly different order. This instruction was emphasized by the fact that the target picture appeared in a randomized position within each trial. In addition, participants were asked to complete the initial two words into a correct sentence of English.

Results The individual answers were manually assessed for correctness, i.e. whether or not it was possible to identify the intended referent from the answer. Out of 192 responses, only 2 ($\approx 1.0\%$) did not allow for a unique identification of the

target picture and were excluded. For the remaining 190 trials, we manually annotated which properties were explicitly mentioned in the response. Figure 4 shows the proportion of properties used in referring expressions by condition. It can be seen that in the pragmatic and simple-size condition both color and size were used in the vast majority of produced referring expressions, while in the simple-color condition only 34% of the referring expressions also contained size.

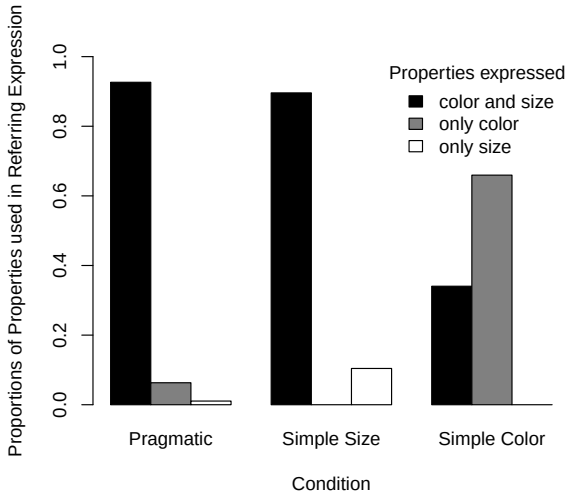

Figure 4: Results of Experiment 1: Proportions of properties expressed in referring expression by condition.

While the difference between the simple-size and simple-color condition has been reported in earlier studies (e.g. Gatt et al., 2011), we observed that in the pragmatic condition speakers use the property color even more often than in simple-size condition (98.9% vs. 89.6%). A logistic regression fit to an indicator of whether color was realized in answers in the pragmatic and simple-size conditions revealed that this difference was significant ($p < .05$).

Discussion These results show that in the production of referring expressions, people tend to express color even if it is not necessary for a successful identification of the intended referent, like in our simple-size and pragmatic conditions. Less salient properties, like size, on the other hand, are used less often if they are not necessary to identify a referent, like in our simple-color condition. These results are a full replication of earlier studies on overspecification in the production of referring expressions (e.g. Gatt et al., 2011). More importanly, we found that in a 'pragmatic' visual context, which would allow for the identification of a referent with just one property through pragmatic reasoning, all but one participant chose to express the property color.

Experiment 2: Comprehension

Experiment 2 is a comprehension experiment, in which listeners are asked to decide which of the possible referents from the pragmatic condition Figure 3 is most likely the intended one given a sentence containing to only one of the referent's two properties. While our comprehension experiment is very similar to the ones reported by Frank and Goodman (2012) together with the model in Equation 1, there are four critical differences: first, Frank and Goodman (2012) asked their participants to place bets on each referent, which were then directly interpreted as (prior or posterior) probabilities for these referents. While this procedure may easily be interpreted as measuring a listener's belief about the likelihood of those three referents, ultimately the listener has to make a decision in order to interpret a given referring expression. We thus opted for a forced-choice task. Second, in line with the previous argument we included the referring expression in a sentence frame. Third, we focused on visual contexts like Figure 3 which require pragmatic reasoning. And fourth, we chose the two properties color and size instead of color, shape and texture used by Frank and Goodman (2012), because for color and size clear differences have been observed in production experiments. In addition, color and size are both typically realized as pre-nominal adjectives, which makes the two corresponding referring expressions only minimally different.

Materials and Design We used the visual stimuli of the pragmatic condition from Experiment 1 (Figure 3), where the target picture shares one property with each of its two competitors: e.g. in Figure 3, it has the same size as the competitor to its right and the same color as the one to its left.

We crossed the two properties of the target referent with the property expressed in the referring expression, yielding a 2x2x2 factorial design of *size* (small vs. large) x *color* (red vs. green) x *referred property* (size vs. color).

We employed a single-shot between-subject design: each participant completed only one trial (cf. Frank & Goodman, 2012). The properties and referring expressions were counter-balanced across participants and the order of the three pictures was randomized within the array.

Participants and Procedure Using Amazon Mechanical Turk, 227 participants were recruited for the experiment. All participants were naive to the purpose of the experiment and reported to be native speakers of English.

The experiment was presented on a web page and consisted of the picture array and a sentence like *Pick the large light bulb.* Participants were asked to select the 'best' picture given the sentence and made their selection by clicking on a radio button under one of the pictures. Like in Experiment 1, participants were instructed to imagine that they are communicating with another person over an instant messaging or chat system and that they both saw the same three pictures, however in possibly different orders, and that they wanted to identify the picture their partner was referring to by the given sentence.

Results All 227 participants selected a picture compatible with the given referring expression. The proportions of listeners' choosing the target referent following from pargmatic reasoning is shown in Figure 5.

It can be seen that if the referred property is size, target choices are fairly similar across all target property conditions, while if the referred property color there is a huge difference in target choices depending on the size of the target: if the target is large, participants mainly selected the target, while for small targets participants had a strong preference for the (large) distractor of the same color. This pattern was confirmed by a logistic regression with the three experimental factors and all their interactions as predictors, which revealed main effects of referred property ($p < .001$) and target size ($p < .001$), and a significant interaction of referred property and target size ($p < .001$).

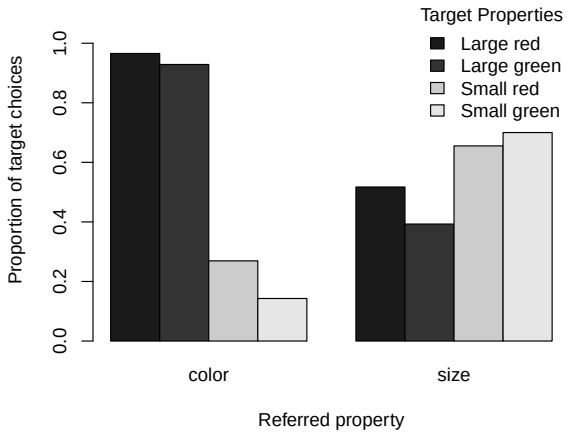

Figure 5: Results of Experiment 2: Proportions of target choices by condition.

Discussion We observed that if the referred property is color, the likelihood of target choice depends strongly on the size of the target (and the size of the distractor of the same color): if the target is large, participants mainly selected the target, while for small targets participants had a strong preference for the (large) distractor of the same color. For size as the referred property, on the other hand, no such difference was observed.

From the perspective of models of pragmatic reasoning, these results are unexpected, as the the visual scenes are symmetric in both properties (i.e. the target shares one property with either of its competitors) and either property should allow target identification through pragmatic reasoning. Instead our results suggest that listeners do not perform much pragmatic reasoning when the referred property is color, but instead follow a simple perceptual heuristic by choosing the most salient referent of that color, i.e. the larger one. This possibility may be accounted for by the salience prior $P(r)$ in the model of Frank and Goodman (2012). In the following section we therefore compare our empirical results to the predictions of this model.

Comparison with Bayesian model

In order to compare our results from Experiment 2 with the predictions of the Bayesian model in Equation 1, we estimated the salience prior $P(r)$ for the three referents in our visual scenes.

Participants, Materials, Design and Procedure 116 participants were recruited for the experiment over Mechanical Turk. The visual stimuli are the same as in Experiment 2, but instead of reading a sentence with areferring expression, participants were told that a speaker had said a sentence they could not understand and were asked to select the picture the speaker was most likely referring to. Like in Experiment 2, we employed a single-shot between-subject design.

Results & Discussion Participants' responses in the salience experiment were used to calculate the prior probabilities $P(r)$. These were combined with the speaker likelihoods from Equation 2 to obtain predictions of the Bayesian model by (Frank & Goodman, 2012). Figure 6 shows the results from Experiment 2 plotted over these model predictions. It can be seen that the model predictions are closer to the observed target choices if the referred property is size than when it is color. This corroborates the speculation that listeners may not perform much pragmatic reasoning when the only property in the referring expression is color.

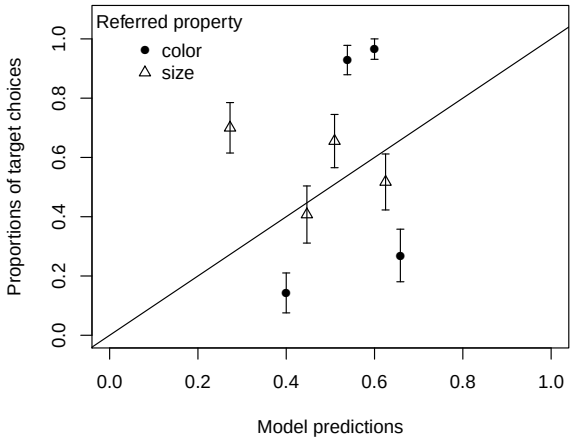

Figure 6: Empirical results over Baysian model predictions by property used in the referring expression. Error bars are standard errors.

General Discussion

The main findings of this paper are the results of Experiment 2: in a language game involving simple pragmatic reasoning, listeners behave differently if the referring property is color from when it is size. In the case of color, target choice is strongly determined by the size of the target: if the target is large, participants mainly selected the target, while for small

targets participants had a strong preference for the (large) distractor of the same color. In the case of size, on the other hand, no such difference was observed. One way to interpret our results is to assume that for referring expressions involving color, listeners do not perform much pragmatic reasoning, but instead follow a simple perceptual heuristic by choosing the most salient referent of that color, i.e. the larger one.

From the perspective of models of pragmatic reasoning, these results are unexpected, even when the salience (prior) of each referent in the context is taken into account, as shown in the preceding section. From the perspective of models of content selection, on the other hand, these results seem rather plausible: as we showed in Experiment 1, speakers often use color to refer to referents, which cannot be further identified by color. While this preference is already very strong if there is a uniquely identifying referent property of lower salience, it is nearly categorical if no such unique property exists and producing a single-property referring expression would involve pragmatic reasoning. As a consequence, listeners may take this speaker preference into account when reasoning about referring expressions. In particular, since color is often used without contributing any discriminatory power, it may be a rational strategy to ignore color as a cue for pragmatic reasoning and instead rely on a simple perceptual heuristic.

References

Baumann, P., Clark, B., & Kaufmann, S. (2014). Overspecification and the cost of pragmatic reasoning about referring expressions. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1898–1903). Austin: Cognitive Science Society.

Benz, A., & Van Rooij, R. (2007). Optimal assertions, and what they implicate: A uniform game theoretic approach. *Topoi*, *26*(1), 63–78.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 2–11). Paris.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gatt, A., Krahmer, E., van Gompel, R., & van Deemter, K. (2013). Production of referring expressions: Preference trumps discrimination. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin: Cognitive Science Society.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In K. van Deemter, A. Gatt, R. van Gompel, & E. Krahmer (Eds.), *Proceedings of the Workshop on Production of Referring Expressions (PRE-CogSci'11).* Boston.

Gatt, A., van Gompel, R., van Deemter, K., & Krahmer, E. (2013). Are we Bayesian referring expression generators? In A. Gatt, R. van Gompel, E. G. Bard, E. Krahmer, & K. van Deemter (Eds.), *Proceedings of the Workshop on Production of Referring Expressions (PRE-CogSci'13).* Berlin.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech Acts.* Academic Press.

Jäger, G. (2011). Game-theoretical pragmatics. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of Logic and Language* (p. 467-491). Amsterdam: Elsevier.

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

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

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

Stiller, A., Goodman, N. D., & Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2134–2139). Boston.

Wittgenstein, L. (1959). *Philosophical Investigations*. Oxford: Blackwell.