University of Pennsylvania Working Papers in Linguistics

Volume 26
Issue 1 *Proceedings of the 43rd Annual Penn Linguistics Conference*

Article 24

10-1-2020

An Experimental Investigation of the Role of Uniqueness and Familiarity in Interpreting Definite Descriptions

Sadhwi Srinivas

Kyle Rawlins

An Experimental Investigation of the Role of Uniqueness and Familiarity in Interpreting Definite Descriptions

Abstract

In this study, we follow a long line of researchers in asking about the precise role of uniqueness and familiarity in the semantics of the English definite article the. We attempt to answer this question experimentally, by observing how definite descriptions behave in contexts where a speaker potentially uses an incorrect description, as in Donnellan’s classic martini scenario, where a speaker incorrectly believes there is a unique referent for their chosen description. In particular, we investigate how hearers interpret definite descriptions in contexts that are systematically manipulated to vary in whether they do or don't contain a unique referent satisfying the description, and whether the referent has or has not been made familiar via previous linguistic mention. Our experimental results reveal that both uniqueness (construed as uniqueness with respect to the common ground between the interlocutors) and familiarity (construed as strong familiarity or anaphoricity) can act as helpful cues to the hearer during the interpretation of a definite description. However, their effects are graded, with the presence of uniqueness leading to greater referential success than the presence of familiarity. We discuss the implications of these results on several existing standard theories of definiteness, and implement a version of the Rational Speech Acts model to help explain the ways in which the observed behavioral data cannot be fully explained on these theories.

An experimental investigation of the role of uniqueness and familiarity in interpreting definite descriptions

Sadhwi Srinivas and Kyle Rawlins*

1 Introduction

In this study, we follow a long line of researchers in asking about the precise role of *uniqueness* and *familiarity* in the semantics of the English definite article *the*. We attempt to answer this question experimentally, by observing how definite descriptions behave in contexts where a speaker potentially uses an incorrect description, as in Donnellan's 1966 classic martini scenario, where a speaker incorrectly believes there is a unique referent for their chosen description. In particular, we investigate how hearers interpret definite descriptions in contexts that are systematically manipulated to vary in whether they do or don't contain a unique referent satisfying the description, and whether the referent has or has not been made familiar via previous linguistic mention.

Our experimental results reveal that both uniqueness (construed as uniqueness with respect to the common ground between the interlocutors) and familiarity (construed as strong familiarity or anaphoricity) can act as helpful cues to the hearer during the interpretation of a definite description. However, their effects are graded, with the presence of uniqueness leading to greater referential success than the presence of familiarity. We discuss the implications of these results on several existing standard theories of definiteness, and implement a version of the Rational Speech Acts model (Frank and Goodman 2012) to help explain the ways in which the observed behavioral data cannot be fully explained on these theories.

The rest of the paper is organized as follows. In the remainder of this section, we introduce some of the prominent standard theories of definiteness in the literature. In Section 2, we describe our experimental design and present an analysis of the resulting data. In Section 3, we describe and test our rational computational model. Section 4 concludes.

1.1 Existing Theories of Definiteness

There are two main points on which standard theories of definiteness diverge.[1] First, theories may differ in how they explain the 'uniqueness' effect that commonly seems to arise with definite descriptions. This effect is exemplified in (1-2) below. (1) is infelicitous when uttered out of the blue because classes typically involve more than one student, but on this assumption it is unclear which student is being talked about. However, no such problem arises with (2), where the superlative definite description picks out the unique tallest student in the class. While traditional uniqueness-based theories can be construed as advocating for uniqueness of the referent within a restricted domain of the physical world (Russell 2005, Evans 1977), there are other researchers who have proposed weaker notions of uniqueness such as uniqueness within the common ground (Roberts 2003), and uniqueness within the narrow discourse context (Kamp 1981, Heim 1982).

(1) # <u>The student in my class</u> came to my office hour.
(2) <u>The tallest student in my class</u> came to my office hour.

Definites are also commonly described as being associated with 'familiarity' effects, originally attributed to Christopherson (1939), and developed in detail in Heim (1982). According to familiarity theories, definite descriptions pick out a 'familiar' entity, or one that has been discussed previously in the discourse— as opposed to introducing a newly mentioned entity. This is often described not in terms of familiar entities per se, but familiar 'discourse referents', as opposed to 'novel' or

*We would like to thank the audiences and reviewers at CiALT-2 (Berlin), CUNY 2019 (CU, Boulder) and PLC-43, Daphna Heller, and the members of the semantics lab at JHU for useful feedback.

[1] The literature on definiteness is larger than we can review in this paper. For recent overviews, including discussion that goes beyond what we term 'standard' theories, see Schwarz 2013, Coppock and Beaver 2015, Ludlow 2018, Aguilar-Guevara et al. 2019, Ahn 2019 (among others).

new referents (Karttunen 1976). (3-4) exemplifies such a familiarity effect. In (3), *the man* refers to the already familiar man introduced in the preceding sentence who the speaker met, and without such a familiar referent, (4) becomes odd. Note that (3) is felicitous even when uttered in a room full of other men, apparently undermining several obvious formulations of a uniqueness requirement, but (4) may also be explained on a uniqueness theory as a failure of uniqueness relative to the world. Once again, theories differ in what notions of familiarity they use to explain the effect in (3). The two most prominent definitions that we will consider are *strong familiarity* or anaphoricity (Heim 1982), and *weak familiarity* or familiarity entailed by existence of the referent (Roberts 2003).

(3) (in a room full of men) I met a man yesterday. The man was wearing a kilt.
(4) (discourse-initial, out of the blue) # The man was wearing a kilt.

Below, we review in more detail each instantiation of uniqueness and familiarity that features prominently in the literature. Traditional neo-Russellian theories of definiteness often propose a strong uniqueness requirement, where the referent identified by the definite description must be unique in some suitably restricted subset of the world. This is a more absolute notion of uniqueness, crucially agnostic of the knowledge (or the lack thereof) of the interlocutors themselves. Roberts (2003) refers to this type of uniqueness as *Semantic Uniqueness*, and we will also henceforth adopt this terminology. Such a construal of uniqueness is able to successfully explain the data in (1)-(2), with the speaker's class representing the domain of interpretation.

Roberts herself proposes a weaker notion of uniqueness, which she terms *Informational Uniqueness*. In simple terms, informational uniqueness is satisfied if the definite description picks out a unique referent within the Common Ground shared between the interlocutors, regardless of whether uniqueness holds with respect to the actual world. Roberts' theory also successfully explains the contrast between (1) and (2) once the reasonable assumption is made that both the speaker and hearer know there is more than one student in the class, as well as the familiarity failure in (4).

The third and weakest notion of uniqueness we consider here is that of *Discourse Uniqueness*, originally due to Heim (1982). Here, uniqueness is required to hold only with respect to the discourse context, where the discourse context is defined narrowly as consisting only of those items that have been explicitly introduced via prior mention or via a very high degree of contextual salience[2]. Requiring only this weakest form of uniqueness successfully explains the contrast in (3-4). It is unclear how semantic/informational uniqueness can straightforwardly account for this example.

The discourse uniqueness that Heim (1982) considers necessary to make a definite description felicitous is not conventionally encoded as a presupposition of the definite article itself (in contrast to other uniqueness approaches). Instead, it emerges as a consequence of a different presupposition associated with the definite article, i.e., that the entity or referent picked out by the description is "strongly familiar" in the discourse context. *Strong Familiarity* is primarily said to arise through an explicit previous mention, as in (3). Under a theory of strong familiarity, the discourse uniqueness is only needed in order to enable successful identification of the antecedent, as the presence of more than one strongly familiar entity satisfying the description would result in an unresolvable ambiguity.

Roberts (2003) accepts that strong familiarity is sufficient to license definiteness, but considers it too strong to be a necessary condition. She offers an alternative, weaker characterization of familiarity—termed as *Weak Familiarity*. Any entity whose existence is entailed by the context is said to satisfy weak familiarity. Under this approach, it is the combination of informational uniqueness and weak familiarity that licenses the use of a definite description. For example, the definite in (5) is felicitous in a context where there is only one pen in the vicinity (and therefore only one pen in the common ground), and where both interlocutors are aware of the existence of the pen but it hasn't been mentioned before.

(5) Can you pass me the the pen?

Finally, there are proposals that deem adequately strong notions of uniqueness and familiarity as independently sufficient conditions for the felicitous use of definite descriptions. We will discuss

[2] Heim does not give an account of what items are considered contextually salient, and what the necessary and sufficient conditions for contextual salience are. Here, we restrict ourselves to previous mention alone.

two such proposals, the first of which is due to Schwarz (2009). Backed by German data which shows differential morphological marking for definites interpreted based on uniqueness *vs.* those interpreted anaphorically, Schwarz proposes that languages can allow reference disambiguation using either of the two strategies independently. For languages like English which do not morphologically distinguish between unique and familiar uses of definites, an ambiguity based analysis is suggested. This means that there are two independent lexical entries for English *the*—one of which encodes a uniqueness presupposition while the other is based on strong familiarity.

Farkas (2002) also advocates for an account that admits both uniqueness and (strong) familiarity as sufficient conditions for the felicitous use of definite descriptions, but differs from Schwarz (2009) in how she implements this idea. Instead of an ambiguity analysis, Farkas proposes a theory of 'Determined Reference', where uniqueness and familiarity are (disjunctively) used as pragmatic cues towards reference disambiguation in definite descriptions. The set of data that is successfully explained under both of these hybrid analyses subsumes the data explained independently by uniqueness theories as well as strong familiarity theories.

1.2 Current Study

In the current study, we employ a behavioral experiment to ask which notions of uniqueness and familiarity (of the ones discussed above) are most relevant when it comes to interpreting definite descriptions in English. Participants took part in a communication game wherein they were presented with descriptions of scenarios, and were asked to imagine themselves as one of the characters in the scenario (i.e., the *hearer*). At the end of each description, participants interpreted a definite description uttered by the character that they were interacting with (i.e., the *speaker*). The stories varied with respect to the status of uniqueness and familiarity of the objects featured in them.

All targets and distractors were at least weakly familiar, as a result of their mere existence in the scene. Additionally, in some situations, familiarity was manipulated to make the target or the distractor strongly familiar. This was done by means of an explicit prior mention involving the entity by either the speaker or the hearer. Similarly, while discourse uniqueness always minimally held of the target referent, scenarios varied in whether or not stronger notions of uniqueness (semantic/informational uniqueness) were satisfied. In order to distinguish between the notions of semantic uniqueness and informational uniqueness, we included scenarios that introduced an explicit mismatch in the knowledge states of the speaker and the hearer, as in the Donnellan (1966) scenarios mentioned above. The set of descriptions that participants interpreted in our experiment did not include any 'weak definites' (Poesio 1994 and following work). We focus only on 'referential' (instead of 'attributive', in the sense of Donnellan 1966) uses of definite descriptions.

2 Behavioral Experiment

2.1 Participants, Materials, Design

We tested 120 participants, recruited using Amazon's Mechanical Turk service in exchange for monetary compensation. Data from one participant was excluded due to technical errors. The experiment was approved by the Johns Hopkins University Institutional Review Board, and participants indicated their consent after reading an information letter.

In each trial, participants were presented with the description of a scenario and asked to imagine themselves as one of the characters in it (the *hearer*). At the end of the trial, participants interpreted a definite description uttered by the character that they were interacting with (the *speaker*). The stories varied with respect to the status of uniqueness and familiarity of the possible referents featured in them. An example trial is shown in Fig (1). In every scenario, two to three objects were featured prominently, e.g., the labeled jar and the unlabeled jar in Fig (1). The description to be interpreted (e.g., "the jar of camphor") was independently known by the speaker and hearer to apply to one or more objects in the scene. Crucially, our trials allowed for mismatches between the speaker's and the hearer's knowledge. For instance, in Fig (1), the participant (or hearer) knew that both the labeled and unlabeled jars contain camphor, making the description applicable to either jar from their perspective. However, relative to the auditor (i.e., the speaker), only the labeled jar could be correctly referred to as “the jar of camphor”.

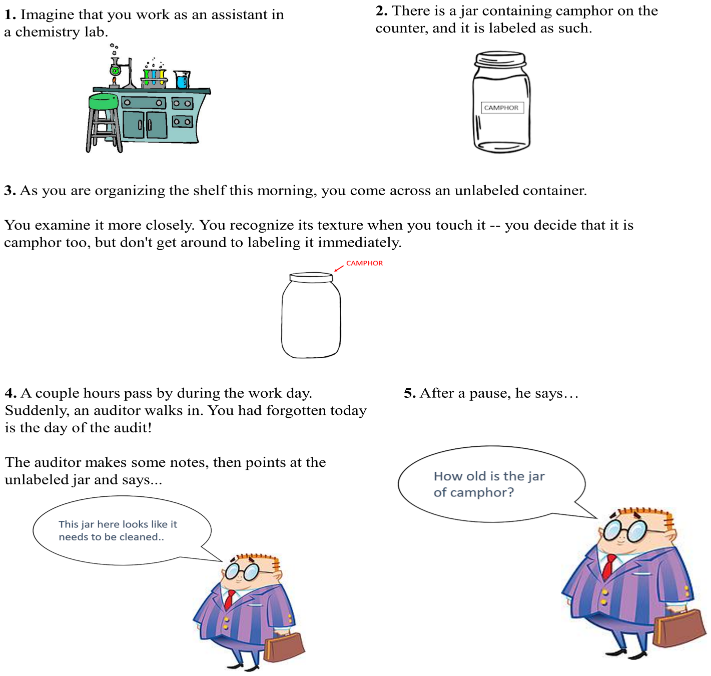
1. Imagine that you work as an assistant in a chemistry lab.
2. There is a jar containing camphor on the counter, and it is labeled as such.
CAMPHOR
3. As you are organizing the shelf this morning, you come across an unlabeled container.
You examine it more closely. You recognize its texture when you touch it -- you decide that it is camphor too, but don't get around to labeling it immediately.
CAMPHOR
4. A couple hours pass by during the work day. Suddenly, an auditor walks in. You had forgotten today is the day of the audit!
The auditor makes some notes, then points at the unlabeled jar and says...
This jar here looks like it needs to be cleaned..
5. After a pause, he says…
How old is the jar of camphor?

Figure 1: An example of a trial where informational uniqueness is true (only the labeled jar is known to be a *jar of camphor* in the common ground) but semantic uniqueness is false. The unlabeled jar is made strongly familiar using a deictic first mention.

We employed a 2 (Semantic uniqueness) x 2 (Informational uniqueness) x 3 (Strong familiarity) within-subjects design, giving rise to 12 within-subjects conditions in total. **Semantic uniqueness** was true within a trial as long as there was a unique referent in the scene that the description could literally apply to. The presence or absence of the semantically unique object was always known to the hearer. **Informational uniqueness** was true if there was a unique referent that the description applied to in the common ground shared between speaker and hearer, regardless of actual facts. In Fig (1) then, informational uniqueness is true because of the presence of a unique jar of camphor in the common ground—i.e., the labeled jar; however, semantic uniqueness is false (as known to the hearer), since both the labeled and unlabeled jars contain camphor.

The final factor that we systematically manipulated was **Strong familiarity**. In any trial, a referent was strongly familiar if it had been mentioned prior to the utterance of the description to be interpreted, either by the speaker or hearer. Every session contained three types of trials with respect to this factor. First, there were trials in which no referent was strongly familiar. Second, there were trials in which the strongly familiar referent satisfied the description as per both speaker and hearer’s knowledge. Finally, there were trials in which the strongly familiar referent did not satisfy the descriptive content as per one or both of the interlocutors. For instance, in Fig (1), the strongly familiar referent is known by the hearer to be a jar of camphor, but not by the speaker.

The description to be interpreted was always of the form *the NP*. However, the form of the first mention (when the referent was strongly familiar) was more variable: they could be definite expressions, indefinite expressions, proper names or deictic expressions. Every subject participated in 12 trials, with each trial exemplifying one of the 12 possible conditions, and the order randomized. To minimize any item effects, we adopted a fully crossed design where there were twelve instantiations of each of twelve different story outlines, such that each instantiation corresponded to a different condition. Each story was seen by ten participants (thus resulting in a total of 120 participants), and no participant was exposed to more than one instance of the same story outline. Participants were instructed to take as much time as they needed. At the end of each trial, participants were asked which object in the scene was intended by the speaker as the referent of the definite description. They were allowed to choose between one of the prominently featured objects in the scene (e.g., either the labeled or the unlabeled jar in Fig 1), or refrain from picking any object by selecting the option *Don't know* instead. In addition, participants could choose to issue a clarification request in the form of a constituent question or a yes/no question, but we will not analyze that data here.

2.2 Results

We performed a mixed-effects logistic regression, where the dependent variable indicated whether participants picked one of the objects in the scene as the intended referent of the definite description ("referential success"), instead of choosing the option *Don't know* ("reference failure"). Our model included semantic uniqueness, informational uniqueness and strong familiarity as categorical fixed factors which were weighted-effects coded. We looked for main effects of each of these factors, as well as interactions between them. The model included random intercepts for participants and story outlines. In addition, we computed random slopes of the fixed factors for every story outline in an attempt to measure any item effects. Below, we first report the results from comparing the effectiveness of semantic *vs.* informational uniqueness as cues towards the interpretation of definite descriptions. Following this, we report the effect of strong familiarity.

2.2.1 Semantic *vs.* Informational Uniqueness

The results indicate that the presence of informational uniqueness is more relevant in interpreting definite descriptions and leads to greater referential success (β = 1.54, p $<$ 0.001), although semantic uniqueness also has a small effect ($\beta = 0.57$, p < 0.001). This suggests that participants in the experiment reasoned with respect to the common ground shared by the speaker, rather than their private beliefs. The data are shown in the red bars in Fig 2. Participants chose a referent about 87.5% of the time when informational uniqueness was true. Of these, 86.2% of all choices corresponded to the informationally unique referent, establishing informational uniqueness as an important clue towards the identity of the intended referent. In the absence of informational uniqueness, a referent was chosen only 46.1% of the time on average. In these cases, the chosen referent was mostly the strongly familiar one, rather than the semantically unique one.

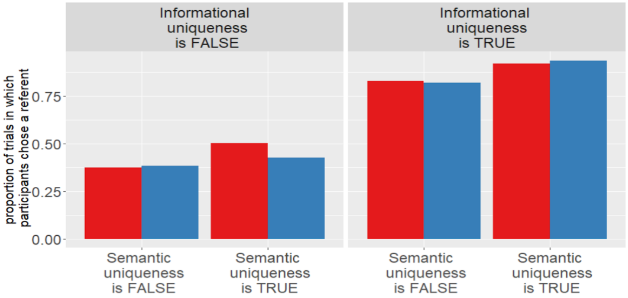

Figure 2: % referential success in varying conditions of informational and semantic uniqueness. The red bars indicate the actual behavioral data obtained in our experiment. The blue bars indicate predictions made by the rational computational model described in Section 3.

2.2.2 Informational Uniqueness *vs.* Strong Familiarity

Due to space constraints, and since the more relevant notion of uniqueness used in interpreting definite descriptions was found above to be informational rather than semantic uniqueness, we will here discuss only the effectiveness of strong familiarity against that of informative uniqueness. The participants' data are depicted in the green bars in Fig (3), and lead to two main observations. First, Fig (3) shows that informational uniqueness is sufficient for interpreting the description, regardless of the status of strong familiarity of any object in the scene.

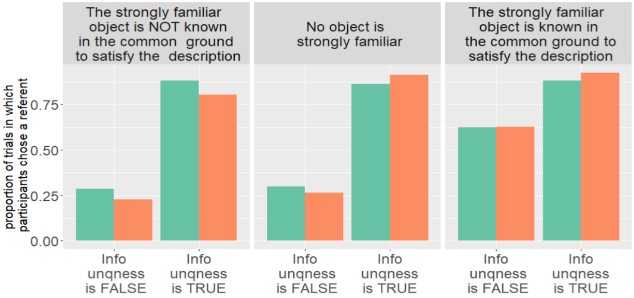

Figure 3: % referential success in varying conditions of uniqueness and strong familiarity. The green bars indicate the actual behavioral data obtained in our experiment. The orange bars indicate predictions made by the computational model described in Section 3.

Second, in trials where informational uniqueness was absent, the presence of a strongly familiar referent was found to lead to referential success about 62.7% of the time, but only when it was known in the common ground that the familiar object satisfied the description. In trials where the previously mentioned object did not satisfy the description as per the common ground, strong familiarity was largely irrelevant. Note that these trials include ones where the strongly familiar object was known by the hearer to satisfy the description but not the speaker. This provides further evidence that the participants reasoned with respect to the common ground, rather than their own private beliefs.

Regression results showed a significant main effect of strong familiarity (β=0.54, p=0.002), however this effect is smaller than the one estimated above for informational uniqueness. We also observed significant interaction between informational uniqueness and strong familiarity (β=-0.59, p<0.001), as well as semantic uniqueness and strong familiarity (β=-0.44, p=0.003).

2.3 Discussion

There are two main insights to be gathered from our experimental results. First, we found that informational uniqueness is used to a greater extent than semantic uniqueness in interpreting referring expressions, showing that participants in the experiment reasoned with respect to the common ground they shared with the speaker, rather than their own private beliefs. Second, we found that when faced with a situation where more than one referent in the common ground satisfies the description but only one is strongly familiar, hearers show a marked tendency to choose the strongly familiar referent. However, strong familiarity is found to be less effective on average than informational uniqueness in leading to referential success.

These results are not exactly anticipated by any existing theory of definiteness in the literature. They provide partial support for Roberts' (2003) theory of informational uniqueness; but this theory does not anticipate the effect of strong familiarity in the absence of such uniqueness. Our results also partially accommodate the strong familiarity theories (Kamp 1981, Heim 1982), but these seem to overestimate the helpful effect of strong familiarity in the absence of uniqueness.

The results are somewhat more compatible with hybrid theories of definiteness that see both uniqueness and familiarity as being important factors affecting the interpretation of definite descriptions. However, these theories still do not by themselves lead us to expect the gradedness that we observe between the effects of uniqueness and familiarity, i.e., that uniqueness is apparently more important, and that familiary operates only as a partially successful backup cue. It is somewhat more problematic to account for such gradedness under an ambiguity theory of the definite determiner, where the null hypothesis is that the two cues will be equally effective, than under an underspecification theory like that of determined reference which does not commit to the exact mechanism by which uniqueness and familiarity act to lead to referential success.

Even though our results are suggestive of an under-determined semantics for the definite determiner in the spirit of Farkas (2002), they can only be treated as a preliminary step towards subsequent experimental investigation. For one thing, while strong familiarity was less effective than uniqueness on average, we still observed substantial variability across trials. There were some trials where strong familiarity helped as much as 90% of the time in achieving referential success, in some other trials it was as low as 20%. This indicates that there might have been pragmatic factors within individual trials that we did not control for—such as the perceptual salience of the objects in the scene prior to the mention, or the precise forms of the first and subsequent mentions—that interacted with strong familiarity. It may turn out that once all pragmatic factors are accounted for, an ambiguity analysis can explain the observed data well enough, but the need for further work is clear.

3 Computational Model of Reference Resolution

The goal of this section is to describe a precise computational model of the process by which participants in our study reasoned to arrive at the intended referent of the definite description. The results hint at an interpretation process that is sensitive (to varying extents) to informational as well as semantic uniqueness, as well the presence of strongly familiar objects in the context.

To explain the main patterns observed in our experimental data, we will build on the Rational Speech Acts (RSA) model for reference resolution described in Frank and Goodman (2012) (a.o.). This model is a natural choice with which to try and explain our data for two main reasons. First, in an experiment such as ours which mimics natural discourse situations in letting the uniqueness and familiarity of referents vary independently of one another, a Bayesian model proves useful since it provides fairly direct ways of operationalizing the notions of uniqueness and familiarity. The likelihood term within the equation for Bayes' rule provides a proxy for encoding uniqueness, while the prior term can be used to model effects of strong familiarity.

Second, the RSA framework explicitly models how a conversational agent accounts for their interlocutor's knowledge state via iterative Bayesian reasoning. This is quite directly useful in our case—given that our data suggest that participants primarily reasoned with respect to the common ground they shared with the speaker, rather than their own private beliefs. We do not intend to suggest that the RSA model is the only one that can accomplish these two desiderata (see e.g., Heller et al. (2016)), but we leave exploration of other models to future work.

In the following subsections, we first describe the computational model in Frank and Goodman (2012). Then, we describe a set of modifications to the original model that equips it to handle the specific format of our experimental data. Finally, we report the effectiveness with which such a model is capable of describing our experimental data.

3.1 Description of the Basic RSA Model

Consider a situation in which a speaker utters a description D_1 to refer to one of two objects contained within the scene: object A and object B. Let us further assume that the intended referent is A, while B is a distractor. The description is chosen from among a finite set of possible descriptions $\{D_1, D_2, D_3, \ldots, D_n\}$. It is now the hearer's job to try and map D_1 to the intended referent A. In effect, we are interested in maximizing $p(A \mid D_1)$, which denotes the probability with which the hearer picks A, upon hearing D_1. The RSA is an iterative model, which assumes that both the speaker and hearer are rational agents who make decisions by reasoning recursively over increasingly sophisticated communicative partners. The simplest instantiation of the RSA is a two-level model where a rational listener RL (whose behavior is of interest to us) reasons about a rational speaker RS, who in turns reasons about a naive or literal listener, LL.

When D_1 is uttered, LL uses the literal semantics of D_1—notated as $[\![D_1]\!]$—in deciding whether to choose A. The precise rule is as in (6). $[\![D_1]\!](A)$ evaluates to 1 just in case object A satisfies the descriptive content of $[\![D_1]\!]$, and to 0 otherwise. $P(A)$ denotes the prior probability of A being chosen in the context, even before the D_1 is uttered. When there is no particular reason to expect that one referent is more likely to be talked about than another before the utterance of a description, the prior probability mass is assumed to be distributed among all the referents equally.

The job of the rational speaker *RS* is to decide whether to use the description D_1 to describe the intended referent *A*, out of the set of possible descriptions $\{D_1, D_2, \ldots, D_n\}$. To do this, they reason over the behavior of *LL* according to (7), in order to try and maximize the chance of successful interpretation of the description while minimizing the cost *C* of the utterance, usually proportional to its length or some other type of complexity. $\alpha > 0$ denotes a rationality parameter—the higher the value of this parameter, higher the rationality of *RS*. Finally, the rational listener *RL* reasons about *RS*'s behavior according to (8) in order to decide which of the two objects *RS* intends to describe. (8) simply restates the standard Bayes' rule: $P_{RS}(D_1|A)$ represents the likelihood that the rational speaker utters D_1 to describe *A*, $P(A)$ is the prior probability of choosing *A*. We assume, as is standard in applications of the RSA model, that the participants in our experiments are instantiations of the rational listener *RL*.

$$(6) \quad P_{LL}(A \mid D_1) \propto [\![D_1]\!](A) \cdot P(A)$$

$$(7) \quad P_{RS}(D_1 \mid A) \propto exp(\alpha(logP_{LL}(A \mid D_1) - C(D_1)))$$

$$(8) \quad P_{RL}(A \mid D_1) \propto P_{RS}(D_1 \mid A) \cdot P(A)$$

3.2 Modifications to the Original Model

Here, we discuss four differences between our experimental conditions and the conditions that have been canonically assumed in previous applications of the RSA. We propose modifications to the original model to handle these differences.

The Set of Possible Utterances $\{D_1, D_2, D_3, \ldots, D_n\}$

Most previous applications of the RSA model have assumed a finite set of hand-crafted utterances, usually in toy contexts, that the speaker must choose from to describe a referent[3]. In our case, where the set of possible descriptions varied from trial to trial, and furthermore each trial potentially allowed for any number of descriptions, we use the following strategy for determining the lexicon.

Consider the same situation as before. To estimate the probability that the speaker would choose D_1 to identify object *A*, we need to also consider alternative expressions that could have been used instead of D_1. Towards this end, we abstract over individual descriptions, and instead group all possible descriptions into the following coarse but exhaustive categories. D_1 may belong to any of these four categories in any trial (reasoning from the point of view of the speaker in our experiment): (i) Descriptions that apply to A only, (ii) Descriptions that apply to B only, (iii) Descriptions that apply to both A and B, (iv) Descriptions that apply to neither A nor B.

Incorporating Interlocutors' Knowledge Mismatch

Typically, the RSA model is employed in situations where the literal semantics are shared by both speaker and hearer. But assuming such a shared semantics wasn't appropriate in our scenarios, which were explicitly designed to provide hearers with privileged knowledge that *wasn't* shared by the speakers. To handle this, we assume that hearers start with a common ground-based semantics (as the experimental results indicated); however, noise could be introduced by the hearer's private knowledge.[4]. This noise measure *hr* was included as a trained parameter in our model. The literal semantics for referents *A* and *B* in the different types of trials computed according to this method are shown in Table (1) below.

Incorporating the Effect of Strong Familiarity

In the general case, when there is no reason to expect that object *A* is more likely to be described the speaker than object *B* (or vice versa), it is reasonable to assume that the prior probability is distributed equally between *A* and *B*. However, once an object is made strongly familiar, it tends to

[3] An exception to this is Monroe and Potts (2015), which circumvents the need for manual specification of the lexicon by inferring the lexicon from a corpus instead.

[4] The choice to treat interference from the hearer's private knowledge as *noise* is entirely a model-internal choice. See Heller et al. (2016) for an alternative view which treats hearers as reasoning independently over both common ground and private beliefs, and using a weighted combination of the two to pick a referent.

Potential targets according to speaker's knowledge	Potential targets according to hearer's knowledge	$[\![D_1]]]\!](A)$	$[\![D_1]\!](B)$
A	A	1	0
A,B	A	1	1 - *hr*
A	A,B	1	0 + *hr*
A,B	A,B	1	1

Table 1: Literal semantics of the definite description D_1 in the various conditions.

be more *salient* in the context when compared to the other weakly familiar items. Such increased salience can be encoded within the model as a redistribution of prior probabilities, such that a greater prior probability mass is assigned to the strongly familiar referent. Intuitively, this corresponds to the idea that items that have been referred to previously are more likely to be referred to again. When referent A is strongly familiar, the updated prior probability values are as shown in Equation (9). The value of the parameter s will be estimated from data.

$$(9) \quad P(A) = 0.5 + s \qquad P(B) = 0.5 - s$$

Choosing the option *Don't know*

Previous experimental studies on the interpretation of definite descriptions have typically adopted a forced choice paradigm where participants must necessarily choose one of the possible items within the context as the intended referent of a description D_1. However, in real life, a third response is possible wherein the hearer refrains from picking any referent and instead asks a clarification question. This response was allowed in our experiment—participants could choose the option *Don't know*. Here, we propose a novel way to deduce the probability with which hearers decide to choose the option *Don't know*. Specifically, we hypothesize that this probability, denoted as $P_{\text{RL}}(\neg(A \vee B) \mid D_1)$ is inversely proportional to the difference between $P(A \mid D_1)$ and $P(B \mid D_1)$. Intuitively, if the hearer believes that D_1 is comparably likely refer to either A or B, their uncertainty would lead them to refrain from picking any referent and ask a clarification question instead. The function that determines $P_{\text{RL}}(\neg(A \vee B) \mid D_1)$ must additionally fulfill the desideratum of being bounded between 0 and 1. In light of this, we choose the functional form shown in Equation 10.

$$(10) \quad P_{RL}(\neg(A \vee B) \mid D_1) = \frac{1}{1+e^{\lambda(|P_{RL}(A|D_1)-P_{RL}(B|D_1)|-c)}}$$

The parameters λ and $c$jointly determine the rate at which $P_{\text{RL}}(\neg(A \vee B) \mid D_1)$ changes in response to unit change in the difference between $P_{\text{RL}}(A \mid D_1)$ and $P_{\text{RL}}B \mid D_1)$. It is possible to estimate λ and c from data, but for the sake of simplicity, we treat them as hyperparameters in our model (λ=4, c=0.3).Once $P_{\text{RL}}(\neg(A \vee B) \mid D_1)$ is computed, the final values $P_{\text{RL-final}}(A \mid D_1)$ and $P_{\text{RL-final}}(B \mid D_1)$ are computed by renormalizing as in (11)-(12).

$$(11) \quad P_{\text{RL-final}}(A \mid D_1) = (1 - P_{\text{RL}}(\neg(A \vee B) \mid D_1))P_{\text{RL}}(A \mid D_1)$$

$$(12) \quad P_{\text{RL-final}}(B \mid D_1) = (1 - P_{\text{RL}}(\neg(A \vee B) \mid D_1))P_{\text{RL}}(B \mid D_1)$$

3.3 Model Evaluation

We trained the salience parameter s and the noise parameter hr on our experimental data, with an objective function that maximized the likelihood of the observed data. As mentioned before, $lambda$ and c are taken to be hyperparameters in the model. So are the cost parameter C—set to 1 for all descriptions, and the rationality parameter α—set to a conservative value of 1.

From Equation (9), it is easy to see that the possible values for s can range between 0 and 0.5. s is equal to 0 in a model that assumes no effect of strong familiarity, while s is equal to the maximum value 0.5 in a model assuming maximal effect of strong familiarity. In the latter case, the model's predictions are expected to resemble those made by traditional familiarity-based theories (Heim 1982, Kamp 1981), in that the strongly familiar item is expected to predominantly be chosen as the intended referent. However, given what we observed in our experiment, we might expect s to lie not at either extremity, but somewhere in between 0 and 0.5. Sure enough, the maximum likelihood estimate of s is found to be an intermediate value of **0.13**.

The noise parameter hr could potentially range between 0 and 1, with lower values indicating that hearers reason with respect to the common ground and higher values indicating that they are more prone to reason with respect to their private beliefs. The best fit estimate of the noise parameter hr was found to be a low value of **0.19**—reflecting what was observed in the behavioral experiment. The predictions made by the trained model are shown in the blue bars in Fig (2), and orange bars in Fig (3). As apparent from these figures, the model fits the experimental data very closely.

4 Conclusion

The experimental results obtained in this study point towards a possible need for refinement of the semantics the definite determiner (at least for English): the results are suggestive of a hybrid method of resolving definiteness that incorporates both uniqueness and familiarity, but not in equal measures. However, we leave open the question of whether this needs a novel lexical entry, or a different approach to reference tracking/pragmatics to account for the item effects. Future work must also investigate the full space of models that can characterize the experimental results, beyond just the RSA-inspired model we have presented here.

References

Aguilar-Guevara, Ana, Julia Pozas Loyo, and Violeta Vázquez-Rojas Maldonado, ed. 2019. *Definiteness Across Languages*. Language Science Press.

Ahn, Dorothy. 2019. THAT Thesis: A Competition Mechanism for Anaphoric Expressions. Ph.D. dissertation, Harvard University.

Christopherson, Paul. 1939. The articles. *Study of their Theory and Use in English* .

Coppock, Elizabeth, and David Beaver. 2015. Definiteness and determinacy. *Linguistics and Philosophy* 38:377–435.

Donnellan, Keith S. 1966. Reference and definite descriptions. *The philosophical review* 75:281–304.

Evans, Gareth. 1977. Pronouns, quantifiers, and relative clauses (i). *Canadian journal of philosophy* 7:467–536.

Farkas, Donka F. 2002. Specificity distinctions. *Journal of semantics* 19:213–243.

Frank, Michael C, and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336:998–998.

Heim, Irene. 1982. The Semantics of Definite and Indefinite Noun Phrases. Ph.D. dissertation, UMass.

Heller, Daphna, Christopher Parisien, and Suzanne Stevenson. 2016. Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition* 149:104 – 120.

Kamp, Hans. 1981. A theory of truth and semantic representation. *Formal semantics-the essential readings* 189–222.

Karttunen, Lauri. 1976. Discourse referents. In *Syntax and Semantics 7: Notes from the Linguistic Underground*, ed. J. D. McCawley, 363–385. Academic Press.

Ludlow, Peter. 2018. Descriptions. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University, fall 2018 edition.

Monroe, Will, and Christopher Potts. 2015. Learning in the rational speech acts model. *CoRR* abs/1510.06807.

Poesio, Massimo. 1994. Weak definites. In *Semantics and Linguistic Theory*, volume 4, 282–299.

Roberts, Craige. 2003. Uniqueness in definite noun phrases. *Linguistics and philosophy* 26:287–350.

Russell, Bertrand. 2005. On denoting. *Mind* 114:873–887.

Schwarz, Florian. 2009. Two Types of Definites in Natural Language. Ph.D. dissertation, UMass Amherst.

Schwarz, Florian. 2013. Two kinds of definites crosslinguistically. *Language and Linguistic Compass* 7:534–559.

Department of Cognitive Science
Johns Hopkins University
Baltimore, MD 21218
sadhwi@jhu.edu
kgr@jhu.edu