

Beware of referential garden paths! The dangerous allure of semantic parses that succeed locally but globally fail

Helena Aparicio, Cornell University, US, haparicio@cornell.edu

Roger Levy, MIT, US, rplevy@mit.edu

Elizabeth Coppock, Boston University, US, eecoppock@gmail.com

A central endeavor in psycholinguistic research has been to determine the processing profile of syntactically ambiguous strings. Previous work investigating syntactic attachment ambiguities has shown that discarding a locally grammatically available, but globally failing, parse is costly. However, little is known about how comprehenders cope with *semantic* parsing ambiguities. Using the case study of scopally ambiguous definite descriptions such as *the rabbit in the big hat*, we examine whether comparable penalties arise for non-lexical semantic ambiguities. In a series of reference resolution tasks, we find dispreference for strings that are globally defined but fail to refer under alternative semantic parses, compared to strings where all readings successfully refer to the same individual. Crucially, this effect is only detectable when the alternative failing reading gives rise to a REFERENTIAL GARDEN PATH, where a dynamic constraint evaluation process temporarily settles on a unique referent before eventually failing. We conclude that failing alternative readings cause dispreference for a definite description, but only when the failing interpretation constitutes a red herring.



1. Introduction

When processing a string of words, listeners are often faced with choice points that can have downstream effects. For instance, when interpreting referential expressions, such as definite descriptions, certain parsing decisions could lead to referential failure, while others might lead to success. In this article, we examine the behavioral consequences of parsing ambiguities in the referential domain.

When interpreting local parsing ambiguities involving definite descriptions, listeners have been shown to discard parses that would result in global reference failure early in the time-course of processing an auditory string. In an eye-tracking study using the Visual World paradigm, Tanenhaus et al. (1995) investigate local syntactic ambiguities such as *Put the apple on the towel in the box*, where the PP *on the towel* can be temporarily parsed as either an NP modifier or as the goal. The authors showed that listeners display an immediate preference for an NP modifier interpretation in visual contexts where the goal interpretation would lead to failure. For example, *on the towel* is immediately interpreted as an NP modifier in contexts with two apples, only one of which is on a towel. Similarly, Chambers et al. (2004) examine the processing of temporarily ambiguous sentences such as *Pour the egg in the bowl over the flower*, where the PP *in the bowl* can be temporarily parsed as an NP modifier or as the goal. The authors found that parsing decisions were informed by world knowledge about the action denoted by the verb, as well as the affordances of the different referential candidates. In particular, the PP was preferentially parsed as an NP modifier in contexts containing two eggs, i.e., contexts where the unmodified definite description *the egg* would fail to refer. Importantly, this preference was only present when the two eggs in the scene were in liquid form (in their own separate containers) –i.e., two eggs that could be the object of a pouring action. No comparable parsing preference was observed when one of the two eggs was in solid form –i.e., an egg that cannot be poured– despite the fact that the linguistic label *egg* continued to be a good descriptor of at least two referents in the display.¹

But, on occasion, listeners entertain parses that eventually ought to be rejected. In previous work on global parsing ambiguities, it has been shown that processing costs are incurred by parses

¹ Evidence that listeners strive to avoid failure when interpreting definite descriptions can be found above and beyond the domain of (temporary) parsing ambiguities. For instance, Chambers et al. (2002) present evidence from an eye-tracking study using a referent identification task in which participants heard instructions, such as *Put the cube inside the can*, while looking at visual scenes containing two cans of different sizes. The authors found that, in such contexts, the description *the can* successfully referred as long as only one of the two potential referents was big enough to fit the cube. These results suggest that participants calculated the definite determiner's uniqueness requirement only with respect to referents that were relevant to the action requested in the instruction, in such a way that would prevent the uniqueness check of the definite determiner from failing. Supporting evidence for this claim comes from the interpretation of positive form gradable relative adjectives embedded in definite descriptions, e.g., *the long rod*. Listeners have been found to set the context-sensitive threshold of the relative adjective in such a way that only one individual in the contextually salient comparison class (e.g., a set of rods of varying length) will satisfy the property, thus fulfilling the uniqueness requirement of the definite determiner (Aparicio et al., 2015; Leffel et al., 2016; Ryskin et al., 2019; Sedivy et al., 1999; Syrett et al., 2009).

that are considered and subsequently discarded, while no such penalties are observed when both parses are successful. Traxler et al. (1998) find that reading times for sentences containing adjunct attachment ambiguities involving relative clauses were faster when the two parses were viable, compared to cases where only one interpretation was plausible, a result that is consistent with the view that suppressing the activation of the implausible analysis is costly. These results were replicated by Swets et al. (2008), using a self-paced reading task.² van Gompel et al. (2001) investigate VP-NP attachment ambiguities and find that for sentences with no initial bias towards either attachment, readers displayed higher processing cost when the semantic information was compatible with one single parse, compared to sentences that were compatible with both parses. In a similar vein, van Gompel et al. (2005) also provide evidence from relative clause attachment ambiguities showing that globally ambiguous sentences are easier to process than their disambiguated counterparts (see also van Gompel et al. (2000), and Clifton Jr and Staub (2008) for an overview).

The evidence reviewed so far suggests that while listeners strive to reject failing parses as early as possible, the process of discarding such parses incurs a cost. While all the studies considered so far test instances of syntactic ambiguity, comparable choice points can also arise during *semantic* parsing. For instance, complex definites such as *the rabbit in the hat* are ambiguous between two readings. On the *absolute reading*, the embedded noun phrase is interpreted just as it would be in isolation, as referring to the unique hat in the context. Alternatively, the uniqueness requirement of the lower definite can be calculated over descriptive content that is richer than the plain descriptive content of the inner definite. In this case, the description *the rabbit in the hat* requires not that there be a unique hat, but a unique rabbit-containing hat. We refer to cases in which the uniqueness check of the inner definite in a nested definite description is relativized to descriptive content from the higher NP as *relative readings* or *Haddock readings*, as they were famously observed by Haddock (1987).

Given previous evidence that rejecting a failing syntactic parse can incur processing cost, it is conceivable that comparable penalties should arise in the interpretation of complex definite descriptions. Following up on this work, here we ask whether the existence of an interpretive path that leads to referential failure causes processing difficulty, leading to dispreference for a string, even when that string is defined under an alternative parse. More precisely: in contexts where the definite description is defined under one reading but undefined under another, does the undefined reading play any role in the hearer's ability to establish a referent for the description?

We consider two hypotheses. Our null hypothesis is that alternative failing parses do not cause dispreference for a definite description, as long as there is a successful interpretation. According

² The authors also report that the ambiguity advantage disappeared when readers were systematically asked post-trial comprehension questions that probed the participants' interpretation of the relative clause (e.g., the sentence *The maid of the princess who scratched herself in public was terribly humiliated* was followed by the comprehension question *Did the maid/princess scratch in public?*)

to the alternative hypothesis we test here, the existence of an interpretive path that leads to referential failure can be a cause for dispreference of a definite description, even when that description is defined under an alternative parse. Our specific alternative hypothesis concerns failing readings on which the dynamic constraint evaluation process temporarily settles on a unique referent before eventually failing, a phenomenon that we label *referential garden paths*; see Section 2. We therefore refer to the alternative hypothesis as the *RGP hypothesis*.

In Section 3, we present results from a reference resolution task supporting the RGP hypothesis. In the experiment, participants are presented with partially masked auditory stimuli and asked to choose a referent from a visual display. The choice of referent reveals how the ambiguous string was resolved. In target trials, the ambiguous string always consists of a Haddock description where the embedded noun phrase is adjectivally modified, as in *the rabbit in the big(ger) ****. In certain displays, on certain resolutions of the ambiguous string, the embedded noun phrase (*the big(ger) ****) succeeds in referring, but the complex noun phrase as a whole does not. This is a kind of situation in which a referential garden path is theoretically predicted. The results from this experiment show asymmetries in the choice of referent within the visual display that precisely mirror the predictions of the hypothesis that referential garden paths are associated with a penalty. Our results, thus, suggest that the process of establishing a referent for an ambiguous definite description is impeded by referential success at a local level with a reading that ultimately fails at a global level. More broadly, as we discuss in Section 4, these results show for the first time that alternative failing semantic analyses can disrupt processing, in line with what previous work has shown for syntactic processing.

2. Formal framework

2.1 Haddock descriptions

Definite descriptions such as *the hat* are generally taken to presuppose uniqueness with respect to the descriptive content (*hat*), in some sense. If there is more than one hat around, then *the hat* is infelicitous (unless one of the hats has become distinctively salient through prior mention or other means).³ Whatever this uniqueness requirement amounts to, exactly, it seems to be relaxed when the description is syntactically embedded inside another DP, as in *the rabbit in the hat*. As Haddock (1987) points out, a description like this is perfectly felicitous in a scenario

³ It is not altogether uncontroversial that definite descriptions come with a uniqueness presupposition; an alternative view is that definites encode discourse familiarity; such a view is espoused, for example, by Heim (1982). According to Schwarz (2009), the definite articles of the languages of the world can be divided into *strong* and *weak*, where strong articles are associated with discourse familiarity and weak articles merely express uniqueness. By Schwarz's diagnostics, English *the* is ambiguous between a uniqueness article and a familiarity article. But, as Beaver and Coppock (2015) discuss, the full range of uses for English *the* can be accommodated under the assumption that it is just a uniqueness article, and that familiarity can be viewed as a species of uniqueness. For present purposes, it suffices to adopt a uniqueness theory, as we do not consider cases involving repeated reference to an object.

with multiple hats, such as the one pictured in **Figure 1**, taken from Haddock’s paper. Speaking pre-theoretically, the uniqueness requirement of the lower definite seems to be calculated over descriptive content that is richer than the plain descriptive content of the inner definite: *the rabbit in the hat* requires not that there be a unique hat, but a unique *rabbit-containing* hat. As mentioned above, we refer to cases in which the uniqueness check of the inner definite in a nested definite description is relativized to descriptive content from the higher NP as *Haddock readings* or *relative readings*. When the inner definite description is interpreted solely with respect to the descriptive content contained therein, the interpretation is an *absolute reading*.

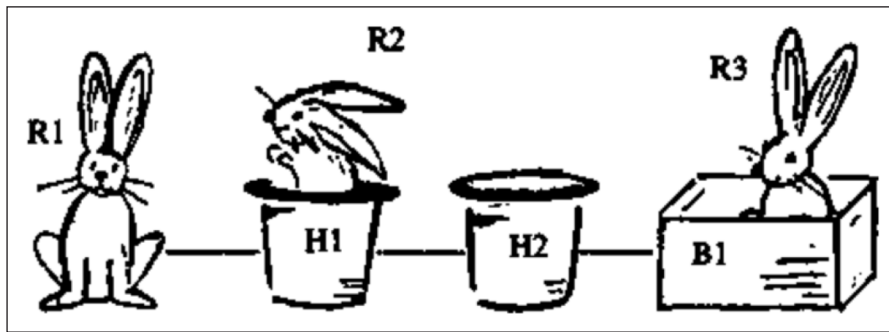


Figure 1: Haddock’s scenario with rabbits, hats and a box.

According to Haddock’s theory, semantic constraints are successively imposed on discourse referents, following the linear order of the words. Haddock envisions a constraint satisfaction problem that can be formalized as a set of open formulas with free variables, like $\text{rabbit}(x)$, $\text{in}(x,y)$, and $\text{hat}(y)$. The uniqueness requirement of the inner definite article is satisfied if there is only one possible satisfier of y left after previous constraints have applied, or, in other words, if the so-called *candidate set* has only one element.

For Haddock, the sequence is determined, for the most part, by the order of the words, with the exception that the definite article’s uniqueness requirement is applied *after* its complement NP’s constraints are incorporated. Each constraint in the sequence has the potential to narrow down on the set of possible values for variables. Setting aside certain details about how variables are unified with each other, the sequence of constraints involved in the interpretation of *the rabbit in the hat* runs as follows, where open logical formulas like $\text{rabbit}(x)$ express constraints. Relative to the scenario depicted in **Figure 1**, the possible values for x and y are updated at each step, as indicated below each constraint.

(1)	Step	1	2	3	4	5
		$\text{rabbit}(x)$	$\text{in}(x,y)$	$\text{hat}(y)$	$\text{unique}(y)$	$\text{unique}(x)$
	x	R1, R2, R3	R2, R3	R2	R2	R2
	y	(any)	H1, B1	H1	H1	H1

As reflected in this example, the application of the uniqueness constraint contributed by each definite determiner is delayed until the corresponding NP is “syntactically closed” (Haddock, 1987, p. 662). The uniqueness requirement of the definite is met if, at the time of application, there is only one remaining possible value for the indicated variable. The theoretical literature has subsequently introduced a number of different mechanisms to derive relative readings along with absolute readings.⁴ In this article, we focus on the account of Bumford (2017), described in further detail below.

2.2 Dynamic interpretation of descriptions

Let us now briefly define a semantic framework in which the notion of referential garden path can be articulated more precisely. In this framework, we will restate Bumford’s (2017) theory of scope ambiguities in Haddock descriptions. This theory delivers precise predictions about when referential garden paths are expected to occur.

We use a version of dynamic semantics in which states are sets of assignments, and formulas determine updates on states so construed. The meaning of a simple definite description in isolation like *the hat* will be represented in this framework as follows:

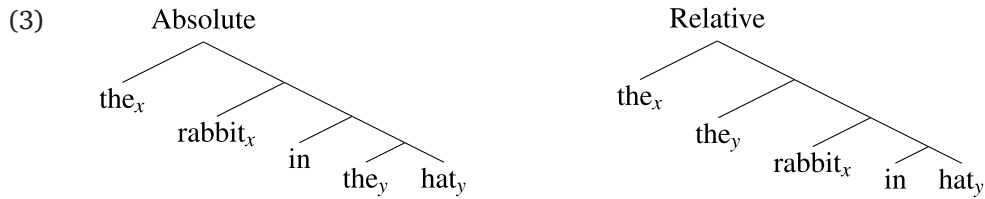
$$(2) \quad \text{hat}(y) + \text{uniq}(y)$$

The effect of the $\text{uniq}(y)$ formula is to ensure that there is only one candidate value for y among the set of values for y provided by assignments in the state.⁵ In other words, the uniqueness check fails, unless there is exactly one candidate value for the discourse referent value when it applies. So, in this case, the uniqueness check requires that there is only one hat. In Appendix A, we define the semantics of formulas like this in a variant of Dynamic Predicate Logic (DPL; Groenendijk & Stokhof, 1989) in the update semantics style presented in Groenendijk and Stokhof (1991).

Bumford (2017) proposes that the meaning of a definite determiner is split between an existential component, which is always interpreted *in situ*, and a uniqueness check, which can either be enforced *in situ* or higher up in the structure of the derivation. The former option, where the uniqueness check is interpreted *in situ*, gives rise to an absolute reading. On the relative reading, the uniqueness check applies at a later stage in the dynamic sequence, corresponding to a higher scope position in the semantic derivation tree. Schematically, the derivation trees for the two readings can be represented as follows:

⁴ van Eijck (1993) uses the existence of relative readings as support for a dynamic theory of meaning (cf. Groenendijk & Stokhof, 1989; Heim, 1983; Kamp & Reyle, 1993), where the interpretation of a noun phrase involves sequential application of constraints on variables. Meier (2003) proposes that embedded definites are predicative and non-presuppositional. Champollion and Sauerland (2010) make use of “intermediate accommodation” to handle it, and Grudzińska and Zawadowski (2019) invoke dependent types.

⁵ It is important that states are sets of assignments rather than single assignments, because the uniq predicate imposes a constraint on the full set of assignments under consideration.



These structures yield different constraint sequences.⁶ The absolute reading involves the sequence of constraints in (4); the relative reading involves the sequence in (5):

(4) $\text{hat}(y) + \text{uniq}(y) + \text{rabbit}(x) + \text{in}(x,y) + \text{uniq}(x)$ (Absolute)

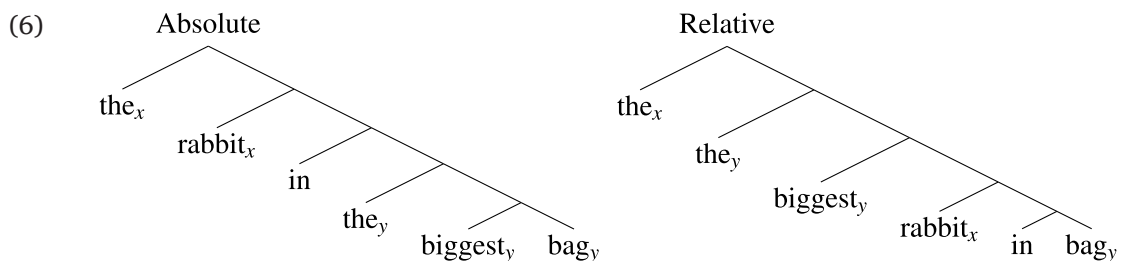
(5) $\text{hat}(y) + \text{rabbit}(x) + \text{in}(x,y) + \text{uniq}(y) + \text{uniq}(x)$ (Relative)

In both cases, each definite article imposes a uniqueness requirement relative to all of the prior constraints that are placed on the associated discourse referent. In the case of the absolute reading, by the time $\text{uniq}(y)$ applies, the only constraint that has been placed on y is that it is a hat. Hence, there must be no more than one hat in total. On the relative reading, uniqueness is checked relative to a richer set of constraints; in this case, there may be more than one hat, as long as there is only one rabbit-containing hat.

2.3 Dynamic interpretation of prenominal modifiers

Our experiments involve nested descriptions containing modifiers in the embedded noun phrase. These expressions provide special insight into the question of how alternative parses play into referential processing, because they can give rise to referential garden paths on certain readings.

Consider, for instance, the nested description *the rabbit in the biggest bag*. This description can have (at least) two interpretations: an absolute interpretation, in which both the definite determiner and the superlative take scope in their surface positions, and a relative interpretation, in which both the definite determiner and the superlative adjective have high scope.



⁶ See Bumford (2017) for the full picture; here, we are glossing over many details of Bumford's analysis.

Now consider the scene depicted in **Figure 2**. Under the absolute reading, the comparison class is all the bags in the scene, and so the embedded description *the biggest bag* refers to the rightmost bag. Under the relative reading, the comparison class is only the rabbit-containing bags, so the embedded description picks out the middle bag, and the nested description as a whole picks out the rabbit in that bag. At the global level, the absolute reading fails relative to the scene in **Figure 2**, because there is no rabbit in the bag that is biggest among all bags. In other words, there is global reference failure under the absolute reading here, because the container that one would describe as *the biggest bag* has no rabbit in it. Although it fails at a global level, there is a local level at which it succeeds; the semantic derivation successfully picks out a unique referent (i.e., the biggest bag, which happens to contain a frog) before crashing. This combination of local success and global failure is what is characteristic of referential garden paths.



Figure 2: A scenario supporting a relative (or Haddock) reading of the complex DP *the rabbit in the biggest bag* but no absolute reading for it.

To model this notion explicitly, let us extend our toolset to account for modifiers. We begin with cases involving superlatives, which Bumford (2017) specifically addressed. In what follows, we review how Bumford’s account works with superlatives, and then extend it to the types of modifiers that we use in our experimental materials, namely, positive form and comparative adjectives.

Loosely following Bumford (2017), we represent the semantic contribution of superlatives with formulas of the form $\text{sup}(v,A)$, where v is some variable and A stands for a gradable adjective. Superlatives filter out assignments to v that are not maximal with respect to A . This treatment implements the idea that the comparison class for the superlative is the set of possible assignment-values for the relevant discourse referent. Note that we are not attempting a compositional semantics here; our goal is only to illustrate the sequence of constraints involved.

The sequence of constraints is determined by hierarchical structure, moving from most embedded to least embedded. According to Bumford (2017), the absolute and relative readings involve the following sequences of constraints, respectively:

$$(7) \quad [[\text{bag}(y) + \text{sup}(y,\text{big}) + \text{uniq}(y)] + \text{rabbit}(x) + \text{in}(x,y)] + \text{uniq}(x) \quad (\text{Absolute})$$

(8) $[\text{bag}(y) + \text{rabbit}(x) + \text{in}(x,y)] + \text{sup}(y,\text{big}) + \text{uniq}(y) + \text{uniq}(x)$ (Relative)

The absolute reading looks for the biggest bag, and then the unique rabbit in it. The relative reading looks for bags with rabbits in them, finds the biggest one, and checks for uniqueness of both the bag and the rabbit at this point.

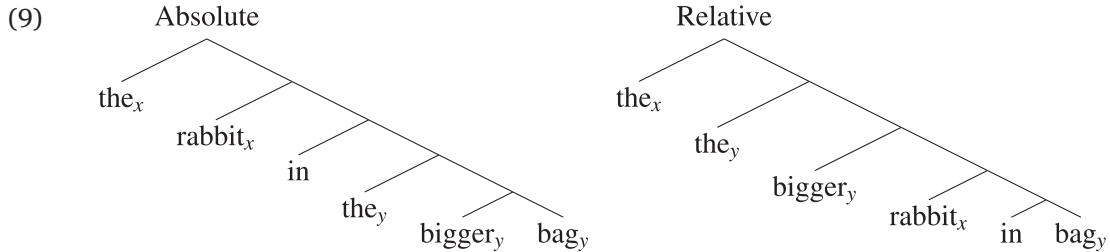
Observe that the superlative modifier is applied after the head noun in both cases, even with the absolute reading. Assuming that a superlative modifier is applied after the noun it modifies accounts for the fact that *the biggest bag* is biggest among bags, and not biggest in general –a fact that Heim (1999) captured by assuming LF movement of the superlative within the nominal. Throughout this article, we assume that modifiers are preferentially applied after the noun they modify, in accordance with the *Head Primacy Principle* of Kamp and Partee (1995, p. 161): “In a modifier-head structure, the head is interpreted relative to the context of the whole constituent, and the modifier is interpreted relative to the local context created from the former context by the interpretation of the head.” The kind of example Kamp and Partee use to motivate their Head Primacy Principle is the contrast between *giant miniature* and *miniature giant*; the first is large for a miniature; the latter is small for a giant. As we have just seen, it has welcome consequences for superlatives as well. It also conveniently cuts in half the number of derivations to consider, and does not affect our qualitative predictions.

2.4 Comparatives

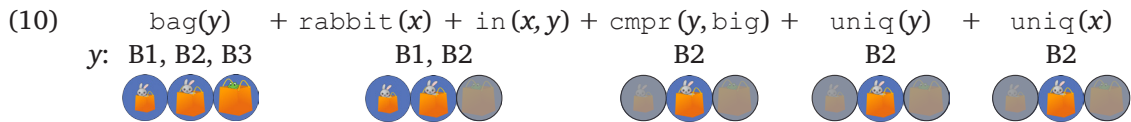
Comparatives exhibit the same kind of relative readings that superlatives have (see Appendix B for a detailed discussion). For a brief illustration, observe that for the three-bag scenario depicted in **Figure 2**, the non-nested description *the bigger bag* is infelicitous, having no identifiable referent. Yet the complex definite *the rabbit in the bigger bag* is perfectly felicitous, picking out the rabbit in the middle bag. It does so on the relative reading, which does refer successfully in this context. To account for both absolute and relative readings of comparatives, we treat comparatives analogously to superlatives. Let us use the formula $\text{cmp}_r(v,A)$ to represent the semantic contribution of a comparative. Here again, A represents a gradable adjective. This formula acts as a filter that keeps only the assignments to v for which there is another assignment in the current state that maps v to something smaller along the A dimension. In a definite comparative like *the bigger bag*, the uniqueness requirement will join forces with the semantics of the comparative to ensure that there are exactly two candidates remaining for the variable in question. This correctly predicts that *the bigger bag* is most felicitous in contexts where there are exactly two bags (see Appendix A for the formal details).⁷

⁷ We acknowledge that our semantics for comparatives is very much made for the special purpose of serving as an attributive modifier of a sortal noun, taking no complements. We do this for the sake of expository simplicity, leaving to the reader’s imagination a more flexible analysis of comparative constructions (covering a wider range of uses) that would give rise to these existential truth conditions for our cases.

We assume that both absolute and relative readings for comparatives exist, just as with superlatives. Schematically, they look as follows:

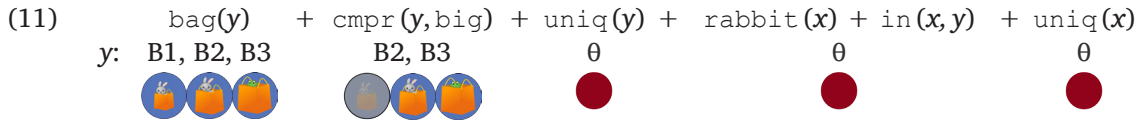


The sequence of constraints corresponding to the relative reading is represented in (10). Below each constraint we depict the set of possible referents remaining for the variable y , relative to the scenario in **Figure 2**, after the constraint has applied (B1 = the smallest bag; B2 = the middle bag; B3 = the biggest bag):



By the time the comparative comes along, the biggest bag has already been removed from consideration. The smallest bag is then ruled out by the comparative, as it is not *bigger* among the candidates. The uniqueness test for y succeeds, then, and the reader may verify that the uniqueness test for x succeeds, as well (the candidates for x are not shown). The middle bag (B2) is the sole candidate for y left by the time all of the constraints have applied, and the sole candidate left for x is the rabbit in the middle bag, in accordance with clear native speaker intuitions about what *the rabbit in the bigger bag* refers to.

The sequence of constraints corresponding to the absolute reading, on the other hand, does not converge on a referent for y (or x):



We use the empty set symbol (\emptyset) to signify that there are no candidates left. When the empty set (visualized by the red dot) begins to appear, it's because a constraint on referents for y failed to be satisfied. Here, the problem is that when the uniqueness constraint applies, there are still two candidates left (the two that count as *bigger*). So the uniqueness constraint fails, and the derivation never recovers. Thus, under the assumptions we have made, relative to the scenario

in **Figure 2**, there is no successful absolute reading for the comparative, although there is a successful relative reading for the comparative.

2.5 Positive form adjectives and referential garden paths

Let us turn now to positive form gradable adjectives like *big*. For positive form gradable adjectives, we assume that the context provides a threshold $\theta(A)$ for each gradable adjective A . A formula of the form $A(v)$, where A is a gradable adjective, filters out assignments of values to the variable v that fail to reach the threshold.

We assume further that, barring failure to assign a referent, only discriminative values of θ are considered. Informally speaking, θ is a *discriminative threshold* for A relative to v in a state σ , i.e., a set of assignments, if θ distinguishes among members of the candidate set so that some count as A and some do not. Hence, for the comparison class in **Figure 2**, two thresholds are possible: one separating the smallest from the two larger bags, and one separating the small and the medium bags from the largest. We call this constraint on thresholds the *non-vacuity principle*, borrowing Kamp and Partee’s (1995) label for the same idea. The non-vacuity principle is defined by Kamp and Partee (1995, p. 161) as follows:

(12) **Non-Vacuity Principle**

In any given context, try to interpret any predicate so that both its positive and negative extension are non-empty.

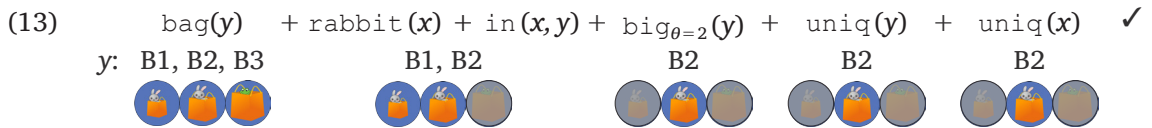
Our framework does not make use of positive and negative extensions, but a similar idea can be captured as a requirement (or preference, rather) that the threshold for a gradable adjective be set in a way that it discriminates among the current candidates for the discourse referent in question.

We understand this “try” as a default mode of comprehension. Evidence that this default can be overridden comes from over-informative uses of adjectives in non-nested descriptions such as *the big blue pin* in contexts with only one pin (Degen et al., 2020). But we take this to be a marked case. Hence, we assume that non-discriminative thresholds are not considered unless they are required in order to identify a referent. In Appendix F, we present evidence in support of the assumption that the non-vacuity principle is only a soft preference for positive form gradable adjectives, rather than a contribution of the semantics. (With comparatives, on the other hand, it is semantically required that there be a non-trivial comparison class.) Another way of thinking about the non-vacuity principle is as a requirement that a gradable adjective be associated with a non-trivial comparison class, along with the assumption that, by default, the comparison class comes from the current discourse context.

Now, moving on to the description *the rabbit in the big bag*, let us consider again the scenario in **Figure 2**. Here, two comparison classes are possible: (i) the set of rabbit-containing bags, so

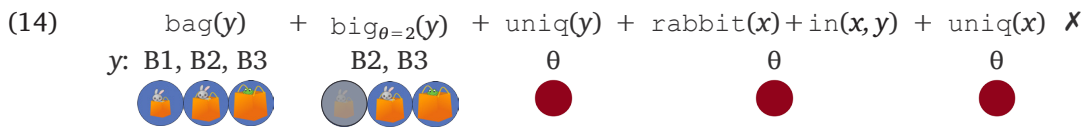
that only the lower threshold is possible; or (ii) the set of all bags, so that either the lower or the upper threshold could be in play. The constitution of the comparison class depends on the order in which the constraints apply.

Let us begin with relative readings. Assume that the smallest bag in **Figure 2** is size 1, the medium bag is size 2, and the largest bag is size 3. As shown in (13), there is a relative reading that succeeds, so long as $\theta = 2$.



If $\theta = 3$, then the gradable adjective applies vacuously, so we assume that such a reading is not considered.

On the absolute reading, the comparison class is just the set of bags, since the only constraint imposed on y by the time the modifier comes along is that it be a bag. Absolute readings fail in the **Figure 2** scenario regardless of whether $\theta = 2$ or $\theta = 3$. They fail in different ways, though. When $\theta = 2$, the uniqueness check for y fails, as shown in (14). In the case where $\theta = 3$, the derivation converges on a referent for y early on, and the uniqueness check succeeds, but ultimately the derivation crashes, as seen in (15). This is the type of situation we refer to as a *referential garden path* (notated \bullet).



In general:

- (16) Relative to a given scene, a linguistic expression gives rise to a *referential garden path* if it has an interpretation on which a unique referent is assigned to some discourse referent at one stage in the dynamic process, but there are no candidates left by the end.

An interpretation (dynamic sequence) assigns a unique referent to a given variable at some stage if there is only one candidate left at that stage, i.e., if the candidate set is a singleton.

We hypothesize that this local success in identifying a referent leads listeners astray, causing a dispreference.

The term *referential garden path* is inspired by the notion of garden path familiar from the parsing literature (Bever, 1970; Ferreira & Henderson, 1991; Frazier & Rayner, 1982; Garnsey et al., 1997; Trueswell et al., 1993). Referential and parsing garden paths differ in at least one obvious way: while parsing garden paths involve alternative syntactic parses, referential garden paths involve alternative dynamic-semantic interpretations. Put differently, comprehenders get *referentially* garden-pathed because they commit to a non-viable semantic analysis due to local convergence on a unique referent under that analysis, not because they commit to the wrong syntactic analysis. In the current work, the relevant connection between these two types of garden paths is that they both involve a local, not global, ambiguity (structural in the case of parsing garden paths, and semantic in the case of referential garden paths). The experiment we report on in the next section suggests that when a string is ambiguous between multiple readings, and one of those readings is associated with a referential garden path in the context, the string is dispreferred.

3. Main experiment: Modified nested definites

The goal of the main experiment to be reported in this article is to determine whether there are behavioral signatures of failing readings of scopally ambiguous descriptions. In particular, we aim to determine whether referential garden paths give rise to penalties.

As discussed in Section 2, referential garden paths constitute only one particular type of failing reading; there are dynamic constraint applications that lead to failure without incurring a referential garden path. In Appendix C, we present results from a reference resolution task parallel to that used in the experiment to be presented in this section, in which we examine unmodified nested definites such as *the rabbit in the bag*. We refer to the study reported in Appendix C as the no-modifiers experiment. The experiment aims to quantify potential penalties of failing readings that do *not* involve referential garden paths. More specifically, we investigate whether definite descriptions that are compatible with a relative interpretation, for which the alternative absolute interpretation fails (without incurring a referential garden path), are penalized compared to cases where the relative and absolute interpretations converge on the same referent. We find that the degree of preference for a complex description is *not* negatively impacted when one of the readings, i.e., the absolute reading, fails. This suggests that a complex definite description such as *the rabbit in the bag* is equally acceptable in contexts that support both relative and absolute interpretations and contexts where only a relative interpretation is available. This finding suggests that there is no penalty (at least in this paradigm) for resolutions of the string for which an absolute reading fails as a result of a violation of the inner definite's uniqueness presupposition.

The current experiment focuses on a more stringent type of failure, namely, failures that give rise to referential garden paths. As shown in Section 2, nested descriptions whose embedded noun phrase contains a positive gradable adjective (e.g., *the rabbit in the big bag*) are theoretically predicted to produce referential garden paths in some contexts. Embedded noun phrases with comparative modifiers (e.g., *the rabbit in the bigger bag*) are also theoretically predicted to give rise to referential garden paths, under certain circumstances. Both of these types of constructions feature in the materials for this experiment.

Our results show a drop in target selection rates precisely in the contexts where referential garden paths are theoretically predicted to arise for both relative and absolute interpretations. Overall, our findings suggest that having to discard a locally successful, but globally failing, interpretation creates a dispreference for a string.

3.1 Methods

3.1.1 Design and materials

The experiment consisted of a reference resolution task in which participants were presented with visual scenes containing five possible referents. Each scene was paired with an auditory instruction that asked participants to click on one of the objects in the display. In experimental trials, the auditory instruction contained a nested definite description, manipulated so that the embedded noun was masked with static noise (*Click on the rabbit in the big/ger ****). Given the visual properties of the display, the masked string was compatible with exactly two possible resolutions, i.e., two possible nouns, e.g., *bag* or *box* (see **Figure 3**). The participants' task was to click on the referent that matched the most likely resolution of the masked auditory instruction in a given scene (e.g., the rabbit in the bag or the rabbit in the box).

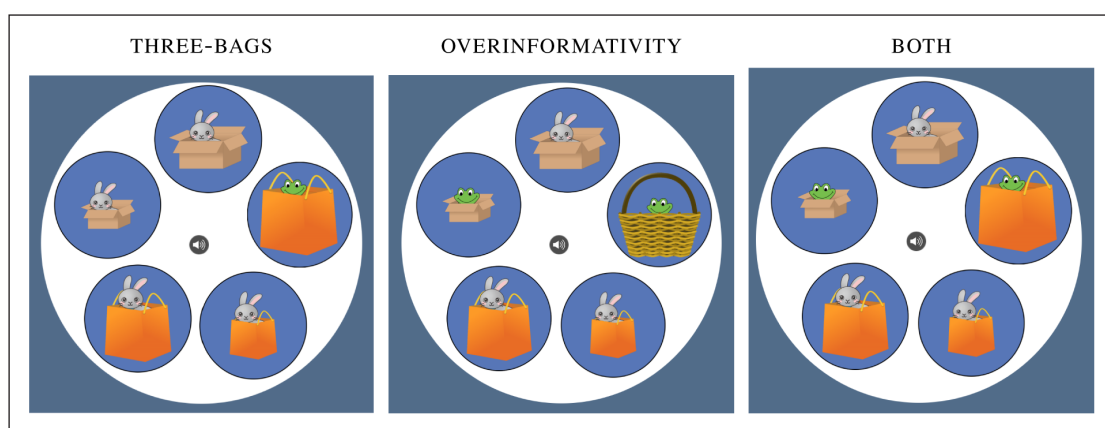


Figure 3: Visual displays tested in the main experiment. Clicking on the audio button in the center of the screen would launch an auditory instruction of the form, “Click on the rabbit in the big/ger ***.”

Experimental stimuli. Twelve items were constructed; see Table 2 in Appendix D for a full listing. Auditory stimuli were recorded by a male native speaker of American English in a soundproof booth. Care was taken to ensure that speech rate, volume, and pitch were as consistent as possible. In experimental trials, the auditory instruction was always of the form ‘Click on [the OUTER-NOUN PREPOSITION *the* MODIFIER INNER-NOUN]’. The inner nouns were masked by replacing the sound with low-amplitude static noise. The INNER-NOUN could, therefore, be realized in two different ways (e.g., *bag/box* for one item, or *pillow/paintbrush* for another). In order to control for coarticulation effects that could provide participants with phonetic cues regarding the identity of the masked noun, the two inner nouns shared an onset (e.g., *bag* and *box*). The modifier was either positive (*big*) or comparative (*bigger*).

Auditory instructions like this were given to participants along with three types of visual scenes, varying along two dimensions. One was the number of objects matching possible resolutions of the inner noun. For instance, in some displays, there were three bags, with the largest containing a frog rather than a rabbit, forcing a relative reading of a string like *the rabbit in the big bag*, and in others, there were only two, allowing for an absolute reading. The other dimension along which scenes varied involved the placement of entities matching the description of the outer noun. For example, in some scenes, there were two boxes, both containing a rabbit, and in other scenes, one box contained a rabbit, and the other contained a frog. This manipulation affected whether or not a modifier would be informative, helping to narrow down on the space of possible referents, or redundant. Although scenes varied along two binary dimensions, we only tested three of the four logically possible combinations (see footnote 12).

In the first scene type, only one of the possible resolutions was compatible with an absolute interpretation. As an example, consider the THREE-BAGS scene in **Figure 3**. Here, Target 1 is a rabbit in a bag, while Target 2 is a rabbit in a box. The scene contains two other rabbits that are inside a smaller bag and a smaller box, respectively. This ensured that the bag and the box associated with the two targets could be felicitously described as *big* (or *bigger*, depending on the instruction). In this case, the *bag* resolution of *the rabbit in the big* *** is not compatible with an absolute interpretation, because the object that would be described as *the big bag* simpliciter –the biggest of the bags– contains a frog. In general, in this type of scene, there is a referent that could not be described using the outer noun in the description, e.g., a frog, and it is associated with the biggest among inanimate objects of the same type as Target 1: a bag, in this case.

Derivations (17) and (18), repeated from (14) and (15), illustrate in detail the fact that the absolute interpretation of the *bag* resolution fails in the THREE-BAGS scene, regardless of the threshold value adopted. Furthermore, there is one derivation corresponding to this resolution that involves a referential garden path.

$$\begin{array}{l}
 (17) \quad \text{bag}(y) + \text{big}_{\theta=2}(y) + \text{uniq}(y) + \text{rabbit}(x) + \text{in}(x,y) + \text{uniq}(x) \quad \times \\
 \quad y: \text{B1, B2, B3} \quad \quad \quad \text{B2, B3} \quad \quad \quad \theta \quad \quad \quad \theta \quad \quad \quad \theta \\
 \quad \quad \quad \text{[img: 3 bags]} \quad \quad \quad \text{[img: 2 bags]} \quad \quad \quad \text{[img: red dot]} \quad \quad \quad \text{[img: red dot]} \quad \quad \quad \text{[img: red dot]} \\
 \\
 (18) \quad \text{bag}(y) + \text{big}_{\theta=3}(y) + \text{uniq}(y) + \text{rabbit}(x) + \text{in}(x,y) + \text{uniq}(x) \quad \bullet^{\times} \\
 \quad y: \text{B1, B2, B3} \quad \quad \quad \text{B3} \quad \quad \quad \text{B3} \quad \quad \quad \theta \quad \quad \quad \theta \\
 \quad \quad \quad \text{[img: 3 bags]} \quad \quad \quad \text{[img: 1 bag]} \quad \quad \quad \text{[img: 1 bag]} \quad \quad \quad \text{[img: red dot]} \quad \quad \quad \text{[img: red dot]}
 \end{array}$$

In (17), which assumes $\theta = 2$, the derivation crashes once the uniqueness check on bags applies. At that point, the candidate set contains two possible referents (B2 and B3), and so the uniqueness check fails. When $\theta = 3$, as in (18), the derivation converges on a single referent for y , i.e., B3. However, upon the application of subsequent filters, the derivation crashes. Therefore, in (18), the failed derivation gives rise to a referential garden path. Given that no value of the threshold variable results in a defined description for the absolute reading of the *bag* resolution, the only possible reading is a relative one. On the other hand, under the *box* resolution, both the relative and the absolute readings are defined. The full range of readings for the THREE-BAGS condition is spelled out in **Figure 4**. We refer the reader to Appendix E for the full set of derivations.⁸




THREE-BAGS scene 				
	Absolute	Relative	Absolute	Relative
<i>big</i> ($\theta = 2$)	\times	\checkmark	<i>big</i> ($\theta = 2$)	\checkmark \checkmark
<i>big</i> ($\theta = 3$)	\bullet^{\times}	\times	<i>big</i> ($\theta = 3$)	\times \times
<i>bigger</i>	\times	\checkmark	<i>bigger</i>	\checkmark \checkmark

Figure 4: Summary of predicted outcomes for modified Haddock descriptions under absolute vs. relative readings in the THREE-BAGS condition. The left panel contains outcomes for the *bag* resolution (Target 1), while the right panel contains outcomes for the *box* resolution (Target 2). \checkmark = success; \times = failure; \bullet^{\times} = failure with referential garden path.

We now turn to the second scene tested, the OVERINFORMATIVITY scene (middle panel of **Figure 3**). In this display, the experimental manipulation involved the informativity of the modifier. The OVERINFORMATIVITY scene was designed such that the adjectival modifier in the instruction was helpfully informative under the *bag* resolution (Target 1), but redundant/over-informative

⁸ In **Figures 4** and **5**, a checkmark signifies only that the description succeeds in referring on the indicated semantic parse; the tables do not reflect informativity violations, which occur with *box* resolutions in the OVERINFORMATIVITY and BOTH scenes.

under the *box* resolution (Target 2).⁹ This was accomplished by replacing the rabbit in the smallest box with a frog.¹⁰

The informativity manipulation causes a referential garden path to arise for the *box* resolution of the comparatively modified description (*the rabbit in the bigger box*) under a relative interpretation (19). Recall that the comparative meaning discussed in 2.4 is a filter that keeps only assignments to v for which there is another assignment that maps v to something smaller. Therefore, once the outer noun filter $\text{rabbit}(x)$ has applied, the output candidate set contains a single referent (namely, the rabbit in the medium bag). However, this referent is later incompatible with the meaning of the comparative, which requires multiple candidates in its input state. When the comparative applies, nothing passes the test. The derivation, thus, yields a referential garden path, because it settles temporarily on one referent and then ultimately fails.

$$(19) \quad \text{box}(y) + \text{rabbit}(x) + \text{in}(x, y) + \text{cmpr}(x, \text{big}) + \text{uniq}(y) + \text{uniq}(x) \quad \bullet^*$$

y : B1, B2 B2 θ θ θ

The full range of (un)available readings associated with the OVERINFORMATIVITY scene are spelled out in **Figure 5**.¹¹ We refer the reader to Appendix E for the full set of derivations.

⁹ We see overinformativity as a property of a given string, rather than as a property of a string on a particular reading. Formally, we may define *overinformativity* as follows:

- (i) A modifier m is *overinformative* in a given description $d = [...m...]$ if there is a description d' such that
- d' is string-identical to d except that m is removed from it
 - d' refers to referent r on all of its successful readings, and
 - d refers to referent r on all of its successful readings.

An interesting case in point is *the rabbit in the bigger box*, relative to a scene with two boxes, one containing a rabbit. The unmodified *the rabbit in the box* succeeds in referring on a relative reading; the modified *the rabbit in the bigger box* succeeds in referring to the same object on an absolute reading. This would constitute a case of overinformativity under our definition.

¹⁰ The inclusion of the prenominal modifier in the THREE-BAGS scene was required for successful reference, regardless of whether the masked portion of the string was resolved as *bag* or *box*; an equivalent but unmodified instruction, such as *the rabbit in the bag/box*, would have failed to refer in the THREE-BAGS scene. This is due to the fact that an unmodified instruction like this could not distinguish between the rabbit in the smallest bag/box and the rabbit in the medium bag/box. Therefore, in the THREE-BAGS condition, the adjective was always helpfully informative.

¹¹ Note that the relative interpretation of the positive adjective under the box resolution would not be considered when $\theta = 2$ is adopted, as such a threshold would not discriminate among referents, i.e., it violates the non-vacuity principle. As discussed in 2.5, we assume that readings involving non-discriminative thresholds are not considered unless they are necessary for establishing a referent. However, if it were to be considered, it would succeed. Therefore, we label this reading with a checkmark.



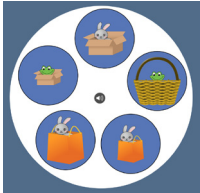
OVERINFORMATIVITY scene					
		Absolute	Relative	Absolute	Relative
<i>big</i> ($\theta = 2$)	✓	✓	<i>big</i> ($\theta = 2$)	✓	✓
<i>big</i> ($\theta = 3$)	✗	✗	<i>big</i> ($\theta = 3$)	✗	✗
<i>bigger</i>	✓	✓	<i>bigger</i>	✓	● ^a

Figure 5: Summary of outcomes for modified Haddock descriptions under absolute vs. relative readings in the OVERINFORMATIVITY scene. The left panel contains outcomes for the *bag* resolution (Target 1), while the right panel contains outcomes for the *box* resolution (Target 2). ✓ = success; ✗ = failure; ●^a = failure with referential garden path.

Finally, the third scene type, called BOTH, combines the number-of-bags manipulation and the informativity manipulation (see right panel of **Figure 3**). For this scene, the predicted readings for the *bag* resolution are just the same as in the THREE-BAGS scene, and the predicted readings for the *box* resolution are just the same as in the OVERINFORMATIVITY scene.¹²

Each of the twelve items had two versions, A and B. In the B versions, the visual displays were constructed with, for example, bags taking the place of boxes and vice versa. In this way, versions A and B differed in which of the two alternative possibilities for the inner noun was the “primary” one. With the Version A stimuli shown in **Figure 3**, *bag* is the primary inner noun and *box* is the secondary one. In the corresponding Version B stimuli, *box* is primary and *bag* is secondary. This was done in order to compensate for any biases that participants may have had toward one specific type of object in the display over the other, independent of our experimental manipulations. We thus had three within-items variables: SCENE (3 levels), ADJECTIVE TYPE (positive vs. comparative), and VERSION (A vs. B). We employed a Latin Square design with $3 \times 2 \times 2 = 12$ lists, ensuring that no participant saw an item in more than one condition, and across the 12 lists, each item appeared in all 12 conditions.

Fillers. Filler items ($n = 24$) consisted of an auditory instruction paired with five images arranged in a circle, constituting a visual display. All fillers were unambiguous, in the sense that the referent could always be determined from the information available in the auditory instruction, given the visual display. Fillers of the first type (10 trials, see **Figure 6**) consisted of auditory instructions that contained simple DPs. The descriptions could either be unmodified (e.g., *the glasses*; 4 fillers), or adjectivally modified, in which case the head noun was masked (e.g., *the shorter ****; 6 fillers). The adjectival modifier was either a color, positive or comparative gradable adjective, and was always globally informative, i.e., reference resolution was not possible without inclusion of the adjective, even after the masked noun was resolved.

¹² The paradigm would have been complete with a fourth condition, involving two boxes, two bags, and a basket, and rabbits in all of the boxes and bags. But we have no reason to expect any deviation from chance in such a condition; any deviation from chance would only have been attributable to noise.

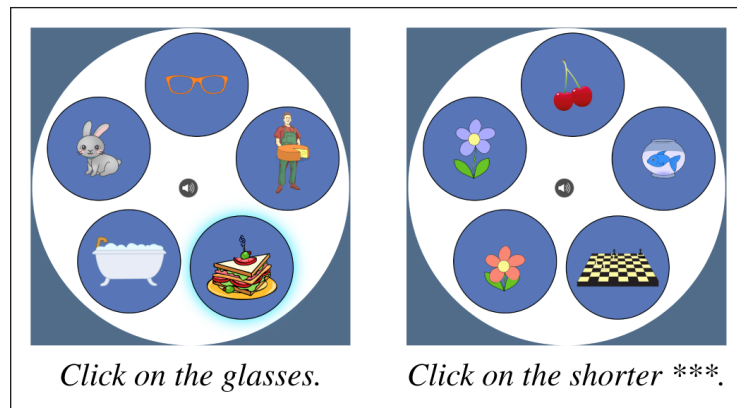


Figure 6: Type 1 fillers: simple definite descriptions.

The remaining 14 fillers contained complex DPs in which one of the NPs was adjectivally modified. As in the first class of fillers, adjectives could consist of color, positive or comparative adjectives. This set of fillers was divided into two main categories, based on whether the inner DP was compatible with an absolute interpretation (7 fillers) or with a relative one (7 fillers). Fillers that could only receive an absolute interpretation were further subdivided into two subtypes: in the first subgroup (4 fillers), the outer noun was masked. This was done in order to counteract the expectation that only embedded nouns would be masked, which helps to keep the participants engaged. The adjective modified the higher NP (e.g., *short *** with the glass*; see the left panel in **Figure 7**) and was always overinformative, since the modifier was not required for the description to successfully refer. To ensure that an absolute interpretation of the inner DP obtained, the visual display contained only one object that could be felicitously described by the inner NP (e.g., a display with one glass; see the left panel in **Figure 7**). In a second subgroup of fillers (3 trials), the inner NP was masked and adjectivally modified. As in the previous subgroup, the adjective was not required for reference identification.¹³ To ensure that an absolute interpretation of the description was possible, the display was designed so that the inner DP could successfully refer in isolation (e.g., a display with two ladders of different size; see the middle panel in **Figure 7**).

In the second main group of fillers (7 trials), the inner NP was masked and the higher NP was adjectivally modified. This was done in order to counteract expectations that adjectivally modified NPs would always be masked. Unlike the first type of fillers, the adjective was necessary for reference identification. The embedded DP could only receive a relative interpretation. This was achieved by including two possible referents for the embedded DP (e.g., two fishbowls; see the right panel of **Figure 7**).¹⁴ Because filler trials were unambiguous, they were also used as attention checks.

¹³ Note that in this case, overinformativity of the adjective was achieved differently than in the target trials, since the higher noun was only compatible with one individual in the display –e.g., a cat, see the middle panel in **Figure 7**–, not two, as in the target trials; therefore, in this case the full PP, including the adjective, was globally overinformative.

¹⁴ A full list of the fillers used in the main experiment can be found in Table 4 in Appendix D.

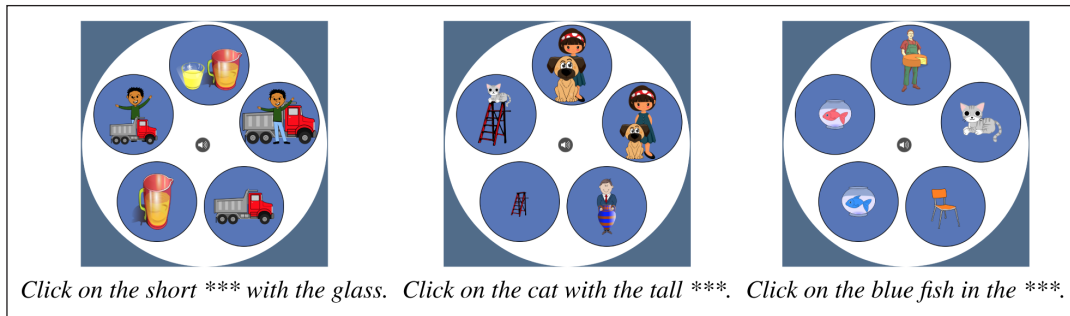


Figure 7: Type 2 fillers: complex definite descriptions.

3.1.2 Participants

We collected data from 242 native speakers of English, recruited through the crowd-sourcing platforms Amazon Mechanical Turk and Prolific. Data from 19 participants was removed from data analysis, due to a failure to pass attention checks, measured as giving more than three unexpected responses in the filler trials. Data from three additional participants was removed, since they took the experiment twice, resulting in a total of 217 participants.

3.1.3 Procedure

The experiment was administered remotely. At the beginning of the experiment, participants were subjected to three practice trials. The practice trials consisted of visual displays of five geometric shapes arranged in a circle. In two out of the three practice trials, the auditory instructions contained a postnominally modified NP (e.g., *Click on the square with stripes*). The third practice trial contained a simple DP with a prenominally modified NP. Two out of the three instructions were masked (i.e., *Click on the red **** and *Click on the triangle with ****). None of the practice trials contained nested definites. In order to trigger the auditory instruction, participants clicked on an audio button in the center of the display. They proceeded to the next trial when they clicked on one of the five images. Clicks to one of the five potential referents triggered the next trial only after the auditory instruction was over. This prevented participants from skipping to the next trial without having heard the full auditory instruction.

3.1.4 Predictions

For this experiment, we tested the specific research hypothesis that semantic parses that give rise to referential garden paths are problematic (the RGP hypothesis). Our linking hypothesis is that resolutions of ambiguous acoustic strings that are associated with problematic semantic parses will be dispreferred, *ceteris paribus*, and that these resolution preferences will be revealed through choice of referent in a visual display. We, thus, expected that our participants' choice of referent would be modulated by two factors: i) whether the associated resolution gives rise to an

informativity violation; and ii) whether the associated resolution has a semantic parse that gives rise to a referential garden path. **Figure 8** summarizes all the referential garden paths associated with each possible resolution in the six experimental conditions.

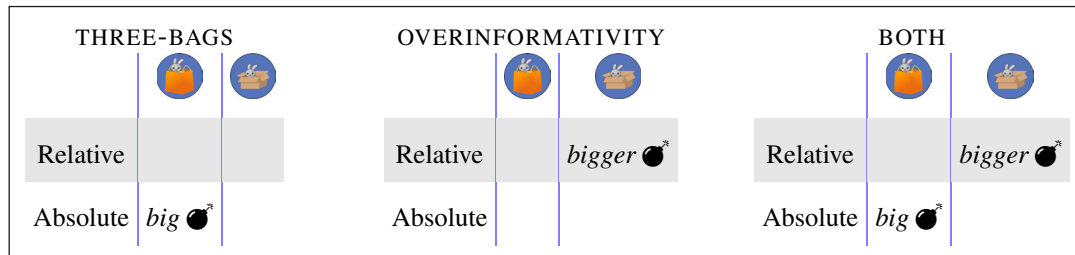


Figure 8: Referential garden paths predicted by the six conditions tested in the main experiment.

Qualitative predictions for each of the conditions tested are summarized in **Figures 9** and **10**. In the **THREE-BAGS** scene, the use of the adjectival modifier is required for successful target identification in both the *bag* and the *box* resolutions. Therefore, informativity should not play a role in this scene's results. The only potential modulator of target selection rates in the **THREE-BAGS** scene should be the referential garden path triggered by the absolute interpretation of the *bag* resolution when the description contains a positive adjective (see the left panel of **Figure 8**). Therefore, under the RGP hypothesis (that failing alternative readings associated with referential garden paths cause dispreference for a string), we should observe lower selection rates for the *bag* resolution over the *box* resolution when the description contains a positive form adjective

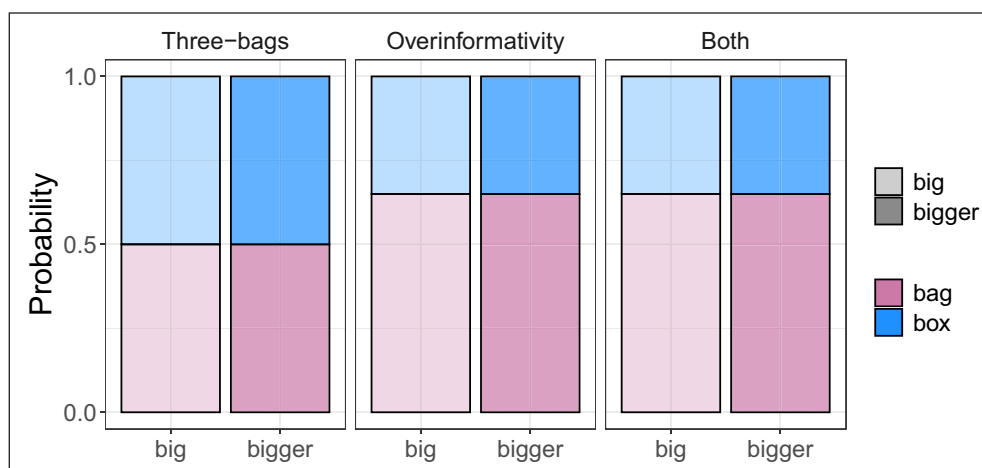


Figure 9: Predictions under the null hypothesis that referential garden paths do not affect referent choices: Participants should be at chance with both *big* and *bigger* in the **THREE-BAGS** scene. There should also be a comparable preference for the *bag* resolution, due to the overinformativity violation associated with the *box* resolution, for the positive and comparative form adjectives in both the **OVERINFORMATIVITY** and **BOTH** scenes.

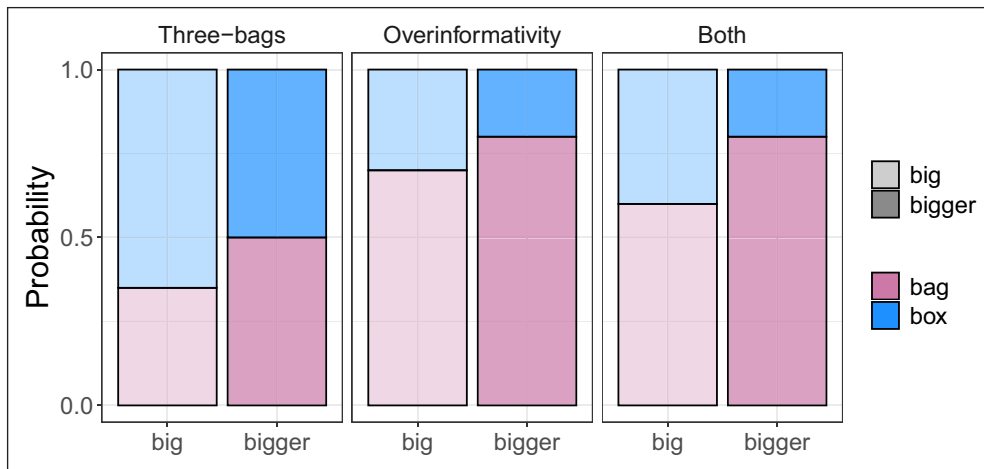


Figure 10: Predictions under the RGP hypothesis (that strings associated with a referential garden path under some reading should be dispreferred, compared to those that do not): For *big*, but not *bigger*, there should be a dispreference for the *bag* resolution in the THREE-BAGS scene, and there should be a difference between the OVERINFORMATIVITY and BOTH scenes for positive form adjectives.

(*big*), such that the target selection rate for the *bag* resolution should be significantly below 50% in this display (see the left panel of **Figure 10**). On the other hand, under the null hypothesis (that failing alternative readings do not have an impact on the degree of preference for a string), both resolutions should be equally probable and participants should be at chance in deciding among them (see the left panel of **Figure 9**). Finally, for descriptions containing a comparative adjective, we predict that the *bag* and the *box* resolutions should be equally probable, since neither involves an overinformative use of the adjective or a referential garden path. Therefore, the two hypotheses under consideration make the same predictions with respect to this condition (see the left panel of **Figures 9 and 10**).

In the OVERINFORMATIVITY scene, the inclusion of a modifier in the instruction is required for successful target identification under the *bag* resolution, whereas the use of the adjective is globally overinformative under the alternative *box* resolution. This should lead to higher-than-chance selection rates for the *bag* resolution, regardless of whether the instruction contains a positive or a comparative adjective, a prediction that is shared by the two hypotheses under consideration. Furthermore, given that the *box* resolution involves a referential garden path under the relative reading with a comparative adjective (see the middle panel in **Figure 8**), the RGP hypothesis predicts that participants should display an even higher preference for the *bag* resolution in the comparative condition, above and beyond the informativity effect, compared to the positive form condition, which does not involve a referential garden path (see the middle panel in **Figure 10**). No such difference is predicted under the null hypothesis (see the middle panel in **Figure 9**).

Finally, let us consider the predictions for the BOTH scene. The makeup of the BOTH scene is a mixture of the THREE-BAGS scene and the OVERINFORMATIVITY scene: the *bag* resolution is identical to that of the THREE-BAGS scene (it contains three bags, but only the two smaller ones contain a rabbit), whereas the *box* resolution mimics the OVERINFORMATIVITY scene (the display contains two boxes: a bigger box containing a rabbit and a smaller box containing a frog). Therefore, we expect our results to display both informativity and referential garden path effects (see the right panel in **Figure 8**). The null hypothesis makes exactly the same predictions in the OVERINFORMATIVITY and BOTH scenes, since no sensitivity to the referential garden path incurred by the absolute reading of the *bag* resolution is expected. The RGP hypothesis, on the other hand, makes different predictions for the BOTH scene compared to the OVERINFORMATIVITY scene. While the results for the comparative are not predicted to be different between those two scenes, the instruction with the positive form should yield a compounded effect of informativity *and* a referential garden path. On top of the informativity penalty for the *box* resolution (common to the OVERINFORMATIVITY and BOTH scenes), we should observe an effect of the referential garden path produced by the absolute reading of the string with the positive adjective under the *bag* resolution. Therefore, with the positive form *big*, we expect that target selection rates for the *bag* resolution should be significantly lower in the BOTH scene, compared to the OVERINFORMATIVITY scene, and higher than the *bag*-selection rates observed in the THREE-BAGS scene (see the right panel in **Figure 10**).¹⁵

3.1.5 Results

Selection rates for the *bag* (Target 1) and the *box* (Target 2) resolutions are shown in **Figure 11** in the six conditions tested. With the exception of the THREE-BAGS scene, participants displayed a preference for the *bag* resolution. Furthermore, within all of the scenes tested, comparative adjectives displayed a higher preference for the *bag* resolution compared to the positive form adjectives. In order to assess whether the predictions of the null hypothesis or the RGP hypothesis are borne out (cf. 3.1.4), we fitted a Bayesian mixed-effects logistic regression model to the entire 2×3 dataset and constructed posterior mean parameter estimates ($\hat{\eta}$ and $\hat{\beta}$, together with 95% symmetric credible intervals (CIs) for each prediction of interest).¹⁶ For ease of interpretation, we

¹⁵ Here, we predict qualitative comparisons between conditions, rather than specific quantitative values, so the visualization in **Figure 10** is more specific than our actual predictions. In particular, we make no prediction about whether target selection rates in the BOTH condition will be above or below chance when the instruction contains the positive form modifier *big*; we predict only that target selection rates for the *bag* resolution should be between those for the THREE-BAGS scene and the OVERINFORMATIVITY scene.

¹⁶ We used the `brm()` function with the default prior in R's `brms` package to fit the model using the maximal random effects structure (random intercepts plus slopes for the two main effects and the interaction, both by-subject and by-item), and used `brms`'s `hypothesis()` function to construct the relevant credible intervals. We ran 6000 iterations and set `adapt_delta=0.9`; for other `brm()` arguments we used the defaults.

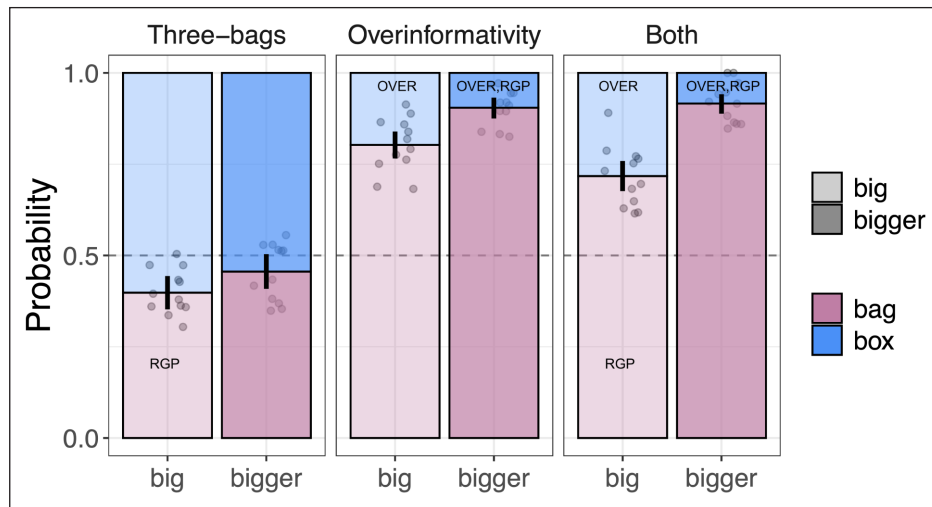


Figure 11: Proportions of responses corresponding to the *bag* (pink) and *box* (blue) resolutions in the main experiment. The error bars represent bootstrap 95% confidence intervals. Floating dots represent item means. Resolutions are labelled with whether they would be overinformative (OVER) or lead to a referential garden path (RGP).

used treatment coding with *big* and THREE-BAGS as the baseline levels for ADJECTIVE and SCENE respectively (for accessible tutorials on treatment coding and other coding schemes, see Schad et al., 2020; Brehm and Alday, 2022). Summary statistics of the fixed effects in the resulting fitted model are shown in **Table 1**.

Let us start with the predictions pertaining to the THREE-BAGS scene. First, do participants disprefer the *bag* resolution, relative to the *box* resolution, in the *big* THREE-BAGS condition, as predicted by the RGP hypothesis? The raw proportions suggest they do (**Figure 8**). This conclusion is confirmed by our statistical analysis: the model’s estimated mean response in this condition is below 50% ($\hat{\eta} = -0.42$, $CI : [-0.65, -0.19]$ in logit space, where a linear predictor of $\alpha < 0$ corresponds to a raw probability of $< 50\%$). In contrast, our model does not predict with 95% confidence a mean response below 50% in the *bigger* THREE-BAGS condition ($\hat{\eta} = -0.19$, $CI : [-0.45, 0.07]$; though note that the model does not quite reach 95% confidence that there is a difference in the mean response between the two conditions: $\hat{\beta} = 0.23$, $CI : [-0.07, 0.54]$).

For the OVERINFORMATIVITY and BOTH scenes, we are interested in assessing the potential effects of three different referential garden paths: one arising from the relative reading of the comparative under the *box* resolution in the OVERINFORMATIVITY scene, one arising from the relative reading of the comparative under the *box* resolution in the BOTH scene, and one arising for the absolute interpretation of the positive form under the *bag* resolution in the BOTH scene (see **Figure 8**). Correspondingly, the RGP hypothesis predicts three simple effects, all of which are visually evident in our data and borne out in our statistical data analysis: an effect of adjective form within the OVERINFORMATIVITY conditions ($\hat{\beta} = 1.11$, $CI : [0.48, 1.94]$), an effect

Table 1: Fixed-effect estimates and 95% symmetric credible intervals for the fitted model. Model specification: $\text{TARGET} \sim \text{ADJECTIVE} * \text{SCENE} + (1 + \text{ADJECTIVE} * \text{SCENE} \mid \text{SUBJECT}) + (1 + \text{ADJECTIVE} * \text{SCENE} \mid \text{ITEM})$. Note that we use treatment coding for both predictors, with the baseline levels being $\text{Adj} = \text{big}$ and $\text{Scene} = \text{THREE-BAGS}$, so that the intercept corresponds to the predicted mean response for the hypothetical average subject and item in the ($\text{Adj} = \text{big}$, $\text{Scene} = \text{THREE-BAGS}$) condition.

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-0.42	0.11	-0.65	-0.20	1.00	15023	10230
$\text{Adj} = \text{bigger}$	0.23	0.15	-0.07	0.53	1.00	17527	9469
$\text{Scene} = \text{OVERINFORMATIVITY}$	2.15	0.23	1.72	2.64	1.00	10136	9661
$\text{Scene} = \text{BOTH}$	1.58	0.23	1.14	2.06	1.00	12244	9163
$\text{Adj} = \text{bigger}:\text{Scene} = \text{OVERINF.}$	0.88	0.39	0.19	1.74	1.00	8253	7776
$\text{Adj} = \text{bigger}:\text{Scene} = \text{BOTH}$	1.68	0.41	0.97	2.59	1.00	7487	7655

of adjective form within the BOTH conditions ($\hat{\beta} = 1.92$, $CI : [1.25, 2.82]$), and an effect of scene within the positive form adjective conditions ($\hat{\beta} = -0.57$, $CI : [-1.12, -0.03]$). In contrast, we do not find clear evidence for an effect of scene within the comparative adjective conditions ($\hat{\beta} = 0.24$, $CI : [-0.85, 1.34]$).¹⁷

3.1.6 Discussion

This experiment has addressed the question of how comprehenders cope with semantic parsing ambiguities. More specifically, our experiment investigated whether the process of discarding a grammatically available, but undefined, semantic parse modulates the degree of preference for scopally ambiguous definite descriptions. Our results suggest a positive answer to this question, but only when the failing parse involves a referential garden path (as under the RGP hypothesis). Our findings show that adult native English-speaking listeners disprefer descriptions that, despite being defined, are associated with referential garden paths under failing alternative parses. The first data point supporting this conclusion is the significant dispreference for the *bag* resolution in the THREE-BAGS scene for the positive form adjective, where the presence of the third, largest

¹⁷ We also tested the 2×2 interaction within this model, which was not non-zero with 95% confidence in this model. However, we found that this result varied depending on the model specification: when we omitted the THREE-BAGS conditions and fit a 2×2 maximal mixed logit model on the remaining four conditions, there was clear evidence for the interaction ($\hat{\beta} = 0.9$, $CI : [0.08, 1.87]$). Likewise, when we reparameterized the 2×3 model so that the baseline SCENE condition was not THREE-BAGS, we found evidence for the interaction at 95% confidence. In contrast, all of the simple effects we report in this paragraph were qualitatively the same, regardless of the model specification. We tentatively speculate that even `brm()`'s default prior specification is not entirely uninformative for the evaluation of this interaction. Regardless, the consistent pattern of simple effects we see is sufficient to clearly affirm our critical theoretical predictions.

bag tempts the listener into an interpretation where the inner definite refers to it. Importantly, in the comparative condition, no such dispreference was detected. This was expected under the RGP hypothesis, as comparative adjectives were not hypothesized to give rise to referential garden paths under any reading in this scene.

The second data point supporting the RGP hypothesis was the significantly lower *bag* resolution rate with the positive form in the BOTH scene compared to the OVERINFORMATIVITY scene. This pattern of results is consistent with the RGP hypothesis, which predicts that because the absolute interpretation of the description containing the positive form leads to a referential garden path under the *bag* resolution in the BOTH scene, but not in the OVERINFORMATIVITY scene, the *bag* resolution should be dispreferred in the former case, compared to the latter. With the comparative modifier *bigger*, on the other hand, the preference for the *bag* resolution did not decrease when a third bag was introduced to the scene. This pattern is as expected, because the *bag* resolution is not associated with a referential garden path for the absolute reading of the comparative description in either of these two scenes.

Finally, referential garden paths not only modulated referent choices pertaining to absolute readings of nested descriptions with positive form gradable adjectives; they also captured subtle effects with comparatives. In particular, a referential garden path was predicted to arise for the relative interpretation of the *box* resolution of the string with a comparative modifier in the OVERINFORMATIVITY and BOTH scenes. No corresponding effect was predicted for the positive form. So the *box* resolution was predicted to be dispreferred in these scenes for comparatives for two reasons –the informativity violation, and the referential garden path– whereas the positive form was only subject to the informativity violation. And indeed, bearing out this prediction, we found that in these two scenes, the baseline dispreference for the *box* resolution was greater for comparatives, compared to positives, as shown by the significantly higher ratings for the *bag* resolution for the comparative, compared to the positive, in both the OVERINFORMATIVITY and BOTH scenes.

The conclusion we draw from this experiment is crucially restricted to one particular type of failing reading, that involving referential garden paths. As mentioned above, in the no-modifiers experiment, reported in Appendix C, we present results from a reference resolution task examining unmodified nested definites such as *the rabbit in the bag*. Such descriptions do *not* involve referential garden paths, and we observe no dispreferences for failing parses in this setting. This finding constrains how generally we can conclude that failing semantic parses are disruptive. Results pertaining to the no-modifiers experiment suggest that a string does not incur a penalty just because it could be given a semantic parse on which it fails to refer; only some failing readings incur a penalty.

To sum up, the current results cannot be accommodated by a theory of semantic parsing ambiguity processing that does not factor in failing alternative parses, at least of the relevant

type. Our findings reveal that defined strings associated with a referential garden path were dispreferred, compared to defined strings that were not.

4 Conclusion

Against the background of the literature on syntactic parsing ambiguities, this article has investigated whether, for global semantic ambiguities, there is a penalty associated with discarding a grammatically viable, but undefined, parse of an otherwise defined string. Our experiments used the case study of scopally ambiguous definite descriptions, such as *the rabbit in the (big/ger) hat*, which are ambiguous between a relative and an absolute reading. Presenting auditory stimuli with partially masked nested descriptions made it possible to assess subtle factors that might affect the preference for one string over another, including the presence of alternative readings that could potentially lead a listener astray. We hypothesized that one factor that might affect this preference is the existence of a *referential garden path*, a failing semantic derivation that converges on a single referent before eventually crashing. The concept of referential garden path yielded strikingly accurate predictions. We detected a dispreference for precisely those resolutions of a string on which referential garden paths were theoretically predicted to arise.

We interpret these findings as evidence that the process of pruning the set of viable semantic parses comes at a cost when the pruned parse constitutes a red herring. Our findings, therefore, converge with previous evidence from the syntactic parsing literature that the rejection of a failing grammatically licensed parse causes processing disruption, a parallel that, to the best of our knowledge, had not been previously established.

Effects of failing alternative readings were *not* detected when such readings did *not* incur a referential garden path. On the contrary, the no-modifiers Experiment (see Appendix C), which tested unmodified Haddock descriptions whose failing readings did not give rise to referential garden paths, yielded a null result. It is, therefore, an open question to what extent it is costly to discard other types of failing analyses. One possibility is that discarding alternative parses is always costly, but failures incurred by a referential garden path are more disruptive compared to other types of failure. This fleeting local success of a parse that is eventually doomed to fail might cause its eventual suppression to be more consequential, resulting in the dispreference for such strings observed in the main experiment.

There is, in fact, some preliminary evidence that the reference failure potential of a string affects its acceptability in a gradient manner. Aparicio et al. (2021) propose that pronominal comparative adjectives such as *the bigger circle* are evaluated against a granularity that maps the individuals in the comparison class (e.g., the set of circles) to degrees in the relevant adjectival scale (e.g., size). The authors argue that pronominal comparatives presuppose that the set of the degrees resulting from this mapping is of cardinality 2. Aparicio and colleagues provide corpus and experimental data showing that definite comparative descriptions are most

frequent and felicitous when evaluated against comparison classes of two individuals, and that their acceptability drops off with higher cardinalities in a gradient manner that is sensitive to granularity. The authors define the *reference failure potential* of a comparative description as the proportion of possible parametrizations of the granularity parameter that map the individuals in the comparison class to more than two degrees, thus causing the description to fail. They find that the reference failure potential of a description is inversely correlated with its acceptability. While not constituting a case of semantic parsing ambiguities, these results suggest that the effect of failing parameterizations was cumulative, at least as measured by the offline acceptability judgment task used by the authors, and that there is a direct relation between the magnitude of the effects on acceptability and the *reference failure potential* of a description. It is, therefore, conceivable that future experiments involving ambiguous referential expressions will detect more subtle penalties for alternative failing readings that do not involve referential garden paths.

The novelty of referential garden paths warrants some discussion about their status, specifically in relation to the better studied garden paths discussed in the parsing literature. It is at this point an open question to what extent syntactic and referential garden paths involve parallel mechanisms. In particular, it is an open question whether the strength of a referential garden path effect is modulated by the likelihood or strength of commitment to the reading that gives rise to it. Our experimental results discourage us from identifying absolute interpretations (or *in situ* scope interpretations) as the default reading in general, since referential garden paths associated with relative readings were also detected in contexts where the absolute interpretation was, in fact, licensed. Rather, it seems that comprehenders entertain *all* possible derivations licensed by the grammar. But the context may shape initial preferences among possible derivations that guide how strongly each is entertained.

In any case, an important takeaway from the current work is the identification of the phenomenon of a referential garden path. The circumstances in which referential garden paths were detected in our study involved different visual scenes, different readings and different types of adjectivally modified descriptions, suggesting that referential garden paths are a general feature of semantic processing. Beyond the experiments presented here, referential garden path effects could potentially arise with other semantic ambiguities involving embedded referential expressions. Looking forward, these findings point to referential garden paths as a new behavioral signature of semantic processing. This has only been a first step in exploring their potential for shedding light on referential communication. Our hope is that referential garden paths will become better understood in future work and prove useful to the psycholinguistics community as a behavioral signature that can be used to probe referential processing as well as the processing of semantic parsing ambiguities.

Appendix A Formal framework extended

Here, we give a variant of Dynamic Predicate Logic (DPL; Groenendijk & Stokhof, 1989) in the update semantics style presented in Groenendijk and Stokhof (1991). The formal language is based on the language of first-order predicate logic, and the semantic value of a formula, given a model, is an update to a given input state. More precisely:

- A model $M = \langle D, I \rangle$ is an ordinary first-order model consisting of a domain of individuals D and an interpretation function I , mapping each non-logical constant (name, predicate, or relation) to its extension in the model.
- An assignment g is a function from variables to elements of D .
- For any term t (individual-denoting variable or constant), $|t|^{M,g} = \begin{cases} g(t) & \text{if } t \text{ is a variable} \\ I(t) & \text{if } t \text{ is a constant} \end{cases}$
- A state σ is a set of assignments.

For a formula ϕ , we write $\sigma \llbracket \phi \rrbracket^M$ to denote the result of updating state σ with ϕ with respect to model M . The superscript M is dropped for readability. The nature of the update depends on the shape of ϕ . As usual in DPL, if ϕ is an atomic formula consisting of a predicate π applied to a sequence of terms t_1, \dots, t_n , then the update keeps assignments that map the terms to tuples in the denotation of π :

- For any predicate π , $\sigma \llbracket \pi(t_1, \dots, t_n) \rrbracket = \sigma \cap \{g \mid \langle |t_1|^g, \dots, |t_n|^g \rangle \in I(\pi)\}$

For conjunction, we use dynamic conjunction, in which the left conjunct applies before the right conjunct, written as $+$.

- $\sigma \llbracket [\phi + \psi] \rrbracket = \sigma \llbracket \phi \rrbracket \llbracket \psi \rrbracket$

We will often omit brackets; please read $+$ as left-associative.¹⁸

Definites. Haddock’s ‘unique’ predicate can only be evaluated relative to the full set of assignments under consideration, because unique expresses a “meta-constraint” (1987, p. 662); in other words, it is a predicate that is evaluated over the whole set of entities to which the relevant variable can be assigned. In dynamic semantics terms, the same idea can be expressed by saying that the update is not distributive, as it does not apply pointwise to each of the

¹⁸ To complete the language, we can make the following assumptions:

- $\sigma \llbracket t_1 = t_2 \rrbracket = \sigma \cap \{g \mid |t_1|^g = |t_2|^g\}$
- $\sigma \llbracket \neg \phi \rrbracket = \sigma \downarrow \phi$,
where $\downarrow \phi = \{g \mid \{g\} \llbracket \phi \rrbracket \neq \emptyset\}$
- $\sigma \llbracket \exists x \phi \rrbracket = \sigma \llbracket \approx_x \rrbracket \llbracket \phi \rrbracket$
where $\sigma \llbracket \approx_x \rrbracket = \bigcup_{g \approx_x h} \{h \mid g \approx_x h\}$

For any two assignments g and h : $g \approx_x h$ means that g and h differ at most with respect to the value they assign to x .

assignments under consideration, but to the whole set of assignments being considered at once. With this in mind, the semantics of Haddock's unique can be defined as follows (dropping the final two letters for typographical reasons).¹⁹

- For any variable v ,

$$\sigma[\![\text{uniq}(v)]\!] = \begin{cases} \sigma & \text{if there is exactly one } k \text{ for which there is a } g \in \sigma \text{ s.t. } |v|^g = k \\ \emptyset & \text{otherwise} \end{cases}$$

Superlatives. We adopt the following denotation for superlatives:

$$(20) \quad \text{For any variable } v, \\ \sigma[\![\text{sup}(v, A)]\!] = \{g \in \sigma \mid \text{for all } h \in \sigma : |v|^g = |v|^h \text{ or } |v|^g >_A |v|^h\}$$

Here, A is a gradable predicate; if $x >_A y$, then the degree to which x is A is greater than the degree to which y is. More formally, we assume that the interpretation function I maps gradable predicates A to individual-degree pairs, so $I(A)$ is a set of pairs $\langle k, d \rangle$ such that individual k has adjectival property A to degree d . Then $x >_A y$ can be defined as: the greatest degree d such that $\langle x, d \rangle \in I(A)$ exceeds the greatest degree d such that $\langle y, d \rangle \in I(A)$. Thus, superlatives filter out assignments to v that are not maximal with respect to A .

Comparatives. For comparatives, we adopt the following definition:

$$(21) \quad \text{For any variable } v, \\ \sigma[\![\text{cmpr}(v, A)]\!] = \{g \in \sigma \mid \text{there is an } h \in \sigma : |v|^g >_A |v|^h\}$$

This lexical entry says that a comparative like *bigger*, interpreted relative to a discourse referent v , acts as a filter that keeps only the assignments to v for which there is another assignment in the current state that maps v to something smaller.

Positive adjectives. The semantic contribution of positive adjectives like *big* is defined as follows.

$$(22) \quad \text{For any variable } v, \\ \sigma[\![A(v)]\!]^\theta = \{g \in \sigma \mid \text{there is a } d \geq \theta(A) : \langle |v|^g, d \rangle \in I(A)\}$$

Non-vacuity. Positive adjectives are also subject to the non-vacuity principle. We define non-vacuity in terms of the notion of *discriminative threshold* as follows:

- (23) θ is a *discriminative threshold* for A relative to v in σ if the set of values for v given by assignments in σ is a proper subset of the set of values for v given by assignments in the result of updating σ with $A(v)$ relative to θ :

$$\{|v|^g \mid g \in \sigma\} \subset \{|v|^g \mid \sigma[\![A(v)]\!]^\theta\}$$

¹⁹ Cf. the 1-operator of Bumford (2017), which is quite similar in meaning to this *uniq*.

Appendix B Relative readings of superlatives and comparatives

It is well-known that superlative adjectives like *biggest* can have relative readings (Bumford, 2017; Heim, 1999; Szabolcsi, 1986; i.a.). One way of seeing this is via the definedness of the description in the lefthand panel of **Figure 12**, where the only available interpretation of the superlative requires the exclusion of bags that do not contain a rabbit, even if they are bigger. The goal of this appendix is to show that comparative adjectives like *bigger* can also give rise to relative readings. The fact that *the rabbit in the bigger bag* is well-defined in the scenario depicted in the righthand side of **Figure 12** is one way of seeing this. Furthermore, as we will show, relative readings of comparatives display all the previously discussed hallmarks of such readings; just like relative superlatives, relative readings of comparatives obviate definiteness effects, are blocked by possessives and by non-modal infinitival clauses, and give rise to similar ambiguities when the adjective modifies the head of a relative clause interpreted in the scope of an attitudinal propositional verb.



Figure 12: Left: A scenario supporting a relative reading of *the rabbit in the biggest bag*. Right: A scenario supporting a relative reading of *the rabbit in the bigger bag*.

As Szabolcsi (1986, i.a.) pointed out, superlatives obviate definiteness effects. In (24a), where *have* takes a complement headed by the relational noun *sister*, the variant with the definite article is quite strange. But there is nothing out of the ordinary about (24b), in which a superlative modifier is added. As Szabolcsi points out (see esp. Sec. 4), examples like (24b) require focus on the subject (here, *Bernie*), and involve a relative reading of the superlative, in that the comparison is among the subject's focus alternatives. Comparatives, too, obviate definiteness effects when they modify a relational noun in the object position of a *have*-sentence.

- (24) a. Bernie has a/??the sister.
 b. Bernie has the nicest sister.
 c. Bernie has the nicer sister.

Just as in the case of the superlative, focus on *Bernie* is required in (24c). Like (24b), it could be used in answer to the question, *Who has the nicer/nicest sister?*, but not *What is Bernie's family like?* (setting aside the relative/intensifier use of *nicest*, in the case of (24b)). Thus, comparatives behave exactly like relative superlatives with respect to this type of definiteness effect.

Second, as discussed by Bumford (2017), relative readings of superlatives are blocked by prenominal possessives:

- (25) a. Who has read the longest play by Shakespeare?
 b. Who has read Shakespeare's longest play? \equiv Who has read *Hamlet*?
Absolute Reading: Who has read the play by Shakespeare that is longer than any other play by Shakespeare?
Missing Relative Reading in (25b): Who has read a longer play by Shakespeare than anyone else has read?

In contrast to (25a), example (25b) can only be interpreted as a question about *Hamlet*, the longest play ever written by Shakespeare. It cannot be construed as asking who read a longer Shakespeare play than anyone else read.

Now let us consider the comparative versions:

- (26) a. Who has read the longer play by Shakespeare?
 b. #Who has read Shakespeare's longer play?
 \rightsquigarrow Shakespeare wrote two plays.
Absolute Reading: Of the two plays written by Shakespeare, who has read the longer one?
Missing Relative Reading (26b): Of the two contextually salient plays written by Shakespeare, who has read the longer one?

Example (26b) implies that Shakespeare only wrote a total of two plays. This is due to a general fact about comparatives in definite descriptions: they tend to be felicitous only when the comparison class contains exactly two elements. Since Shakespeare wrote more than two plays in his lifetime, the absolute reading leads to presupposition failure. In (26a), a relative reading is available, so the sentence is fine; in (26b), only an absolute reading is available, due to the presence of the possessive, so the sentence is forced to carry a false presupposition.

A third parallel between superlative and comparative descriptions can be observed through a phenomenon discussed by Bhatt (2006). Bhatt points out that superlative descriptions associate with focus, giving rise to truth-conditionally different readings depending on the placement of the focused constituent, as shown in (27a) and (27b). Even though Bhatt does not cast the discussion in terms of the relative vs. absolute distinction, we point out that focus-sensitivity effects are contingent on a relative interpretation of the superlative. With an absolute interpretation, the different information structures exemplified in (27a) and (27b) result in non-equivalent truth-conditions, as shown in the associated paraphrases.

- (27) a. Joan_F gave Mary the most expensive telescope.
 b. Joan gave Mary_F the most expensive telescope.
Relative reading for (27a): Some people gave Mary telescopes. Of all those telescopes, the telescope that was given by Joan was the most expensive one.

Relative reading for (27b): Joan gave some people telescopes. Of all those telescopes, the telescope Joan gave to Mary was the most expensive one.

Absolute reading: Joan gave Mary something, and that thing was the most expensive telescope among all the telescopes.

Bhatt observes that association with focus effects are blocked by non-modal infinitival relative clauses (28).

(28) Joan_F gave Mary the most expensive telescope to be built in the 9th century.

Absolute reading: Of all the telescopes built in the 9th century,

Joan gave Mary the most expensive one.

Missing relative reading: Some people gave Mary telescopes built in the 9th century.

Of all those telescopes, the telescope that was given by Joan was the most expensive.

Bhatt's observation can be taken as evidence that non-modal infinitival relatives block relative readings of superlatives.

Examples (29–30) show that comparative adjectives pattern with superlatives in this respect.

(29) a. Joan_F gave Mary the more expensive telescope.

b. Joan gave Mary_F the more expensive telescope.

Relative reading for (29a): Two people gave Mary telescopes. Of all those telescopes, the telescope that was given by Joan was the more expensive one.

Relative reading for (29b): Joan gave two people telescopes. Of those telescopes, the telescope Joan gave to Mary was the more expensive one.

Absolute reading: Joan gave Mary a telescope. It was the more expensive of the two telescopes.

(30) Joan_F gave Mary the more expensive telescope to be built in the 9th century.

Absolute reading: Of the two telescopes built in the 9th century,

Joan gave Mary the most expensive one.

Missing relative reading: Some people gave Mary telescopes built in the 9th century.

Of the two 9th century telescopes she received, the telescope that was given by Joan was the more expensive one.

A final piece of evidence comes from ambiguities such as (31b), again due to Bhatt (2002). Bhatt observes that superlatives modifying a noun heading a relative clause can be interpreted within the scope of a propositional attitude verb inside the relative clause. Example (31c) shows that comparative adjectives exhibit the same ambiguity.

(31) a. the long book that John said Tolstoy had written

b. the longest book that John said Tolstoy had written

c. the longer book that John said Tolstoy had written

High reading: of the books John said Tolstoy wrote, the long/longer/longest one

Low reading: the book John said was long/longer/longest among the ones written by Tolstoy.

On the high reading, both the superlative and the comparative adjective are interpreted against the comparison class of the books mentioned by John, whereas on the low interpretation, the relevant comparison class comprises the books written by Tolstoy. The ordinary positive form gradable adjective *long* has only a high reading, whereas the comparative and superlative adjectives have low readings, as well.

All of these diagnostics show that comparatives have relative readings, just like superlatives. If the Bumford-style analysis of relative readings is right for superlatives, then the analogous kind of analysis should be right for comparatives. In particular, we assume that comparatives can act as filters whose application is delayed in the dynamic sequence, effectively operating at a higher scope position.

Appendix C No-modifiers experiment

The goal of the no-modifiers experiment is to quantify potential penalties of failing absolute readings that do not involve referential garden paths. More specifically, we investigate whether definite descriptions that are compatible with a relative interpretation, for which the alternative absolute interpretation fails (without incurring a referential garden path), are penalized, compared to cases where the relative and absolute interpretations converge on the same referent. The current experiment provides a negative answer to this question. It does so by comparing contexts that support both a relative and an absolute interpretation of an unmodified description against contexts that only support an absolute interpretation of the definite description. Our results show that the degree of preference for a description is not negatively impacted when one of the readings, i.e., the absolute reading, fails. This finding sets up an important point of contrast with respect to the main experiment reported in this article, which shows that *some* alternative readings that fail *do* yield a penalty, namely, those that involve referential garden paths.

C.1 Methods

C.1.1 Design and materials

The experiment consisted of a reference resolution task in which participants were presented with visual scenes containing five possible referents. Each scene was paired with an auditory instruction that asked participants to click on one of the objects in the display. In experimental trials, the auditory instruction contained a nested definite description, manipulated so that the embedded noun was masked with static noise (*Click on the rabbit in the ****). Given the visual properties of the display, the masked string was compatible with exactly two possible resolutions,

i.e., two possible nouns, e.g., *bag* or *box* (see **Figure 13**). The participants' task was to click on the referent that matched the most likely resolution of the masked auditory instruction in a given scene (e.g., the rabbit in the bag or the rabbit in the box).

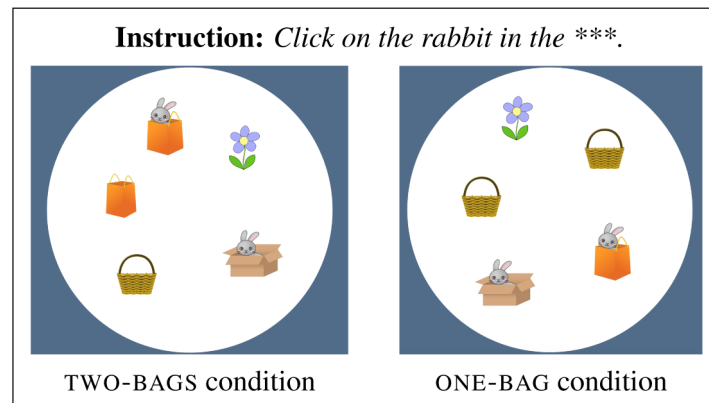


Figure 13: Example scenes for the no-modifiers experiment. **Left panel** represents the TWO-BAGS condition, containing the following five objects: *Target 1*: rabbit in a bag; *Target 2*: rabbit in a box; *Competitor*: empty bag; *Distractor 1*: empty basket; *Distractor 2*: flower. **Right panel** represents the ONE-BAG condition and differs from the TWO-BAGS condition only in that the *Competitor* object is replaced by *Distractor 3*: a second empty basket.

Experimental stimuli. Twelve items were constructed; see Table 3 in Appendix D for a full listing. Auditory stimuli were recorded by a male native speaker of American English in a soundproof booth. Care was taken to ensure that speech rate, volume, and pitch were as consistent as possible. The auditory instruction was always of the form ‘Click on the [OUTER-NOUN] [PREPOSITION] the [INNER-NOUN]’. The inner nouns were masked by replacing the sound with low-amplitude static noise. Given the properties of the visual scene, the masked INNER-NOUN could always be resolved in two different ways (e.g., *bag/box* for one item, or *pillow/paintbrush* for another). In order to control for coarticulation effects that could provide participants with phonetic cues regarding the identity of the masked noun, the two inner nouns shared an onset (e.g., *bag* and *box*). For each of the twelve items, two versions were recorded (Version A and Version B), differing in the choice of inner noun.

Visual scenes were constructed to accompany the auditory instruction. In experimental trials, participants were presented with a display showing five images arranged in a circle, equidistant from each other and from an audio button in the center (see **Figure 13**). The order of the images around the circle on the display was randomly generated for each participant on each trial. Each scene contained two target referents, Target 1 and Target 2, both matching the outer noun in the instruction (e.g., *rabbit*), and realizing one or the other alternative for the inner noun (e.g., *bag* or *box*). The preposition in the instruction (*in* or *with*) determined the spatial relation between the two nouns. As with the auditory instructions, each of the twelve items had two versions, A and B.

Thus, versions differed only in which of the two alternative possibilities for the inner noun was the “primary” one. For example, in the sample stimuli shown in **Figure 13**, *bag* is the primary inner noun and *box* is the secondary one; these are Version A stimuli. In the corresponding Version B stimuli, *box* is primary and *bag* is secondary, so in the B versions, the visual displays are constructed with bags taking the place of boxes and vice versa.

Displays varied according to whether or not an absolute reading was available for the possible resolutions. The main factor of interest was scene type, which had two levels, TWO-BAGS and ONE-BAG. In the ONE-BAG condition, the two potential target referents were compatible with an absolute *and* a relative interpretation of the description (see the right panel of **Figure 13**). In the TWO-BAGS condition, a relative reading for one of the resolutions was enforced by including a third referent, the Competitor, which matched the primary inner-noun alternative (e.g., *bag*). Thus, in the TWO-BAGS condition, Target 2 (e.g., a rabbit in a box) was compatible with an absolute interpretation of the description (i.e., the box resolution in the left panel of **Figure 13**), whereas Target 1 (e.g., a rabbit in a bag) was only compatible with a relative interpretation (i.e., the left panel of **Figure 13**). Depending on the condition (TWO-BAGS or ONE-BAG), two or three additional referents were included as distractors. Distractors could not be described by either the outer noun or the inner noun (e.g., a basket and a flower, or a car and a boat). Finally, the experimental items were distributed in four lists, following a Latin Square design, ensuring that no participant saw an item in more than one condition.

Fillers. A total of 24 fillers were included. Filler trials were of four different types (see Table 5 in Appendix D for the exhaustive list of fillers). The first class of fillers (6 filler trials) contained auditory instructions with simple DPs (e.g., *the bird*, see the left panel of **Figure 14**). In all instances, the description was paired with a visual display that satisfied the uniqueness requirement of the definite determiner. The second class of fillers (6 filler trials) consisted of auditory instructions that contained complex DPs (e.g., *Click on the policewoman in the car*). However, unlike the experimental trials, none of the nouns in the instruction were masked. In 4 out of these 6 filler trials, the embedded DP necessitated a relative interpretation. The third type of fillers also contained complex DP’s in the auditory instruction, and resembled the target items in that the second NP was masked (5 fillers). However, unlike the experimental trials, the masked noun allowed only one possible resolution. The nested DP could either receive a relative interpretation (3 fillers, see panel A of **Figure 15**) or be compatible with an absolute interpretation (2 fillers, see panel B of **Figure 15**). Finally, the fourth class of fillers consisted of complex descriptions where the outer noun was masked (e.g., *Click on the *** with the fish*). PPs in this class varied according to whether the PP could receive a relative interpretation (3 fillers, see panel C of **Figure 15**) or not (4 fillers, see panel D of **Figure 15**). The order of the experimental trials and the fillers was randomized, and experimental trials were pseudo-randomly interspersed with filler trials, ensuring that no more than two fillers were presented in a row. As in the main experiment, filler trials were used as attention checks.

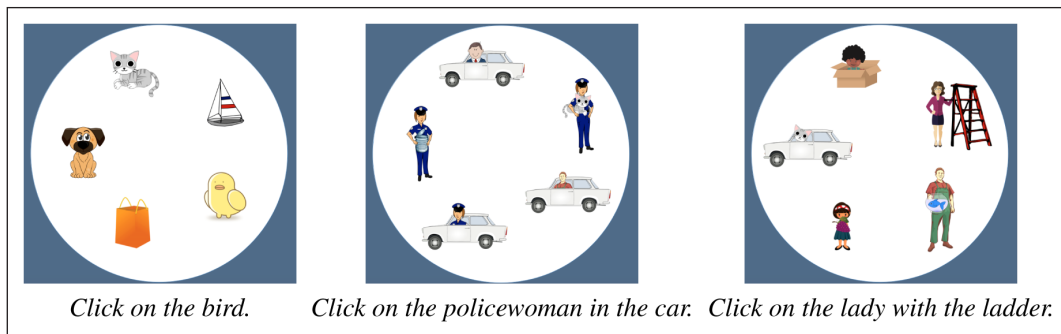


Figure 14: Type 1-2 fillers.

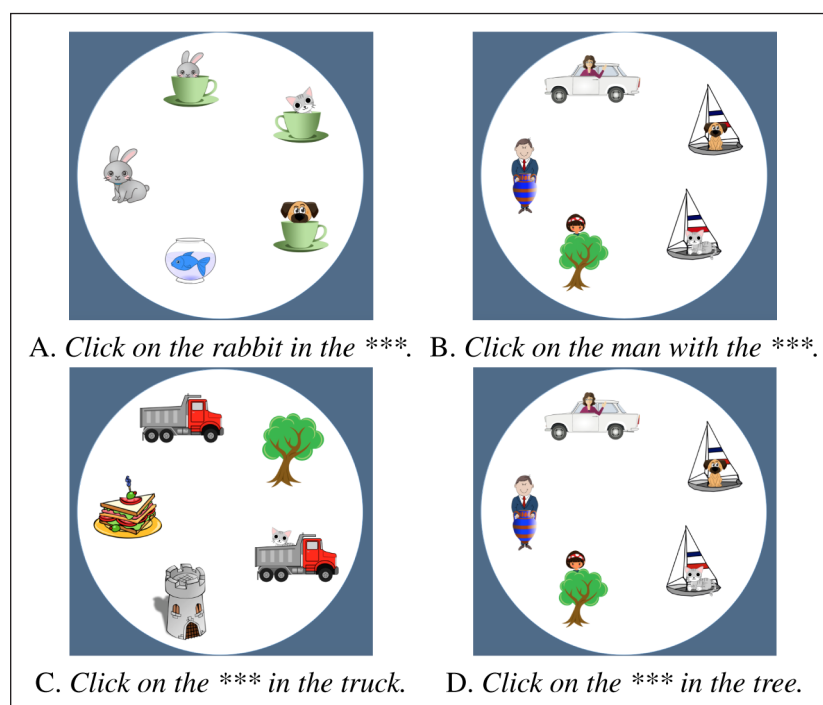


Figure 15: Type 3-4 fillers.

C.1.2 Participants

A total of 52 participants were recruited through Amazon Mechanical Turk and Prolific. Results from 8 participants were discarded, either because they failed to pass our attention check (i.e., giving more than three unexpected responses on filler items), or because they did not self-report being native speakers of English (3 participants). All analyses reported below are based on data from the remaining 41 participants.

C.1.3 Procedure

The procedure was the same as that of the experiment reported in the main text.

C.2 Predictions

As in the main experiment, our linking hypothesis is that given an acoustically ambiguous string, listeners will prefer resolutions of that string that are not associated with problematic semantic parses, and will reveal these preferences through choice of referent in a display. The specific research hypothesis that we are testing in this experiment is that any semantic parse that fails is problematic. This *failure-is-costly hypothesis*, as we will call it, leads to the prediction that strings associated with both absolute and relative readings should be dispreferred when one of these readings fails to converge on a referent.

By comparing target selection rates to the two potential referents, it is possible to gauge participants' interpretive preferences as a function of the available linguistic input and the properties of the visual display. Recall that in the TWO-BAGS condition, no absolute reading is available for the *bag* resolution. Only a relative reading is available; since in this condition there are two bags, the uniqueness check fails for the inner definite under an absolute interpretation. If the failure of an absolute reading causes a dispreference for a given resolution, then it is predicted that there should be fewer clicks to the target object corresponding to the *bag* resolution in the TWO-BAGS condition, compared to the ONE-BAG condition.

C.3 Results and discussion

Figure 16 plots the rate at which participants selected the target objects corresponding to the *bag* and the *box* resolution in the ONE-BAG and TWO-BAGS conditions, respectively. Participants overwhelmingly selected one of the two potential targets, with the exception of one single trial (0.23% of trials). Data pertaining to this trial was removed from data analysis. As observed in **Figure 16**, participants chose the *bag* resolution at chance in both the ONE-BAG and TWO-BAGS

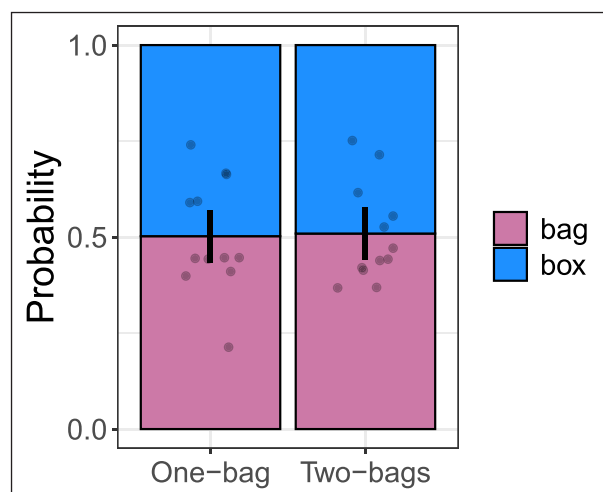


Figure 16: Results of the no-modifiers experiment. Error bars show 95% bootstrap confidence intervals of the mean. Floating dots represent item means.

conditions. Clicks to the target object corresponding to the *bag* resolution (the resolution with the primary inner noun) were submitted to a logistic Bayesian mixed-effects regression model, using SCENE as a fixed effect predictor, and participants and items as random effects. Results did not reveal any significant effects of SCENE ($\beta = -0.2$, $CI = [-0.49, 0.53]$).

The current results establish that a nested definite description, such as *the rabbit in the bag*, is equally acceptable in contexts that support both relative and absolute interpretations and contexts where only a relative interpretation is available. The current results, therefore, contravene the predictions of the failure-is-costly hypothesis, suggesting that there is no penalty (at least in this paradigm) for resolutions of the string for which an absolute reading fails as a result of a violation of the inner definite’s uniqueness presupposition. In other words, comprehenders do not avoid a description just because it could be given a semantic parse on which it fails to refer.

Appendix D Experimental materials

The target items for the main experiment and the no-modifiers experiment are listed in **Tables 2 and 3**, respectively.

Table 2: Target items used in the main experiment.

filler	outer noun 1	outer noun 2	prep.	inner noun 1	inner noun 2	distractor
1	rabbit	frog	in	box	bag	basket
2	frog	bird	in	bathhtub	bucket	boat
3	cat	bird	in	truck	tree	tower
4	boy	girl	with	pillow	paintbrush	pen
5	lady	man	with	fan	fish	flower
6	girl	man	with	dog	duck	doll
7	man	girl	with	sock	scarf	sandwich
8	bird	rabbit	in	can	cup	car
9	monkey	frog	with	glue	grapes	glasses
10	farmer	lady	with	cheese	chair	cherries
11	policewoman	boy	with	carrot	cookie	cake
12	horse	cat	with	ladder	lizard	lamp

Table 3: Target items for the no-modifiers experiment.

filler	outer noun	prep.	inner noun 1	inner noun 2	distractor 1	distractor 2
1	rabbit	in	box	bag	basket	flower
2	frog	in	bathhtub	bucket	boat	duck
3	cat	in	truck	tree	tower	sandwich
4	boy	with	pillow	paintbrush	pen	car
5	lady	with	fan	fish	flower	glue
6	girl	with	dog	doll	duck	cherries
7	man	with	socks	scarf	sandwich	basket
8	bird	in	can	cup	car	boat
9	elephant	with	glasses	grapes	glue	tower
10	farmer	with	cheese	chair	chessboard	pen
11	policewoman	with	vacuum	violin	vase	lamp
12	horse	with	ladder	letter	lamp	vase

The fillers for the main experiment and the no-modifiers experiment are given in **Tables 4** and **5**, respectively.

Table 4: Fillers included in the main experiment.

number	phrase	image1	image2	image3	image4	image5
1	the glasses	big_glasses	rabbit	big_sandwich	farmer-med_cheese	med_bathtub
2	the green ***	green_frog	brown_frog	violin	cat-med_truck	cat-small_truck
3	the blue fish in the ***	med_fish	red_fish	med_chair	farmer-med_cheese	gray_cat
4	the yellow ***	yellow_bird	girl-grapes	elephant	big_cherries	boy-box

(Contd.)

number	phrase	image1	image2	image3	image4	image5
5	the gray *** in the car	gray_cat	rabbit-big_car	cat-med_tree	big_doll	boy-med_paintbrush
6	the cheese	med_cheese	big_tower	med_tree	dog-big_cup	dog-small_cup
7	the tall ***	tall_man	short_man	big_glasses	fork	med_grapes
8	the tall tower with the ***	tall_tower-flag	short_tower-flag	elephant-glasses	bird-med_can	elephant-fan
9	the cat with the tall ***	cat-big_ladder	small_ladder	girl-big_dog	girl-med_dog	man-vase
10	the tall *** with the man	tall_policewoman-man	tall_policewoman	med_truck	med_sock	rabbit-big_car
11	the girl	girl	rabbit-big_box	rabbit-small_box	frog-big_bag	cat-cup
12	the short ***	short_woman	tall_woman	med_pillow	big_sandwich	violin
13	the short pencil with the ***	short_pencil-notepad	long_pencil-notepad	farmer-med_chair	farmer-big_cheese	farmer-small_cheese
14	the dog with the short ***	dog-short_lamp	big_lamp	big_flower	short_flower	man-vacuum
15	the short *** with the glass	short_pitcher-glass	tall_pitcher	man-big_truck	man-small_truck	med_truck
16	the car	big_car	elephant-box	girl-big_pillow	girl-truck	lady-med_fan
17	the taller ***	big_tree	small_tree	big_cherries	fork	med_fish
18	the taller building with the ***	tall_building-antenna	short_building-antenna	man-vacuum	man-med_scarf	horse

(Contd.)

number	phrase	image1	image2	image3	image4	image5
19	the car with the tall ***	car-tall_trafficlight	short_trafficlight	frog-small_bathtub	frog-big_bathtub	frog-med_bucket
20	the tall *** with the dog	tall_boy-dog	short_boy	big_ladder	girl-grapes	big_basket
21	the shorter ***	short_flower	big_flower	big_cherries	med_fish	chessboard
22	the short glass with the ***	short_glass-plate	tall_glass-plate	short_pencil-notepad	long_pencil-notepad	small_ladder
23	the cat with the shorter ***	cat-short_bottle	tall_bottle	tall_boy-dog	short_boy	girl-grapes
24	the shorter *** with the horse	short_farmer-horse	tall_farmer-horse	tall_tower-flag	short_tower-flag	big_tower

Table 5: Fillers included in the no-modifiers experiment.

number	type	phrase	image1	image2	image3	image4	image5
1	*** P the N1	*** in the bathtub	rabbit-bathtub	cat-bag	basket	basket	flower
2	*** P the N1	*** in the box	frog-box	bag	boat	boat	flower
3	*** P the N1	*** in the truck	cat-truck	tower	truck	tree	sandwich
4	N2 P the ***	girl with the ***	boy-pillow	girl-pillow	pillow	paint brush	car
5	*** P the N3	*** with the fish	elephant	lady	lady-fish	farmer	man
6	*** P the N1	*** in the tree	girl-tree	girl-car	girl-truck	tower	tree

(Contd.)

number	type	phrase	image1	image2	image3	image4	image5
7	N1 P the N1	man in the truck	man-truck	girl-truck	cat-truck	truck	truck
8	N2	bird	dog	bird	cat	boat	bag
9	N1 P the N1	elephant with the ladder	elephant-ladder	ladder	ladder	elephant-letter	letter
10	N2	farmer	man	farmer	lady	girl	fish
11	N1 P the N1	police woman in the car	police woman-car	police woman-cat	man-car	farmer-car	police woman-can
12	N1 P the N1	horse with the scarf	horse-scarf	elephant-box	cat-bathtub	fish	rabbit-car
13	N1 P the ***	rabbit in the ***	rabbit-cup	cat-cup	fish	rabbit	dog-cup
14	N3	frog	fish	fan	frog	flower	farmer
15	N1 P the ***	cat in the ***	cat-car	can	cat	car	cup
16	N3	boy	bathtub	boat	boy	box	basket
17	N3 P the N3	lady with the ladder	boy-box	farmer-fish	lady-ladder	girl-grapes	cat-car
18	N3 P the N3	girl with the grapes	girl-glue	girl	girl-grapes	girl-glasses	grapes
19	N2 P the ***	man with the ***	girl-tree	man-vase	lady-car	cat-boat	dog-boat
20	N2 P the ***	bird in the ***	dog-boat	bird-boat	cat-boat	frog-boat	cat-boat
21	*** P the N3	*** with the fan	horse-scarf	elephant-box	elephant-fan	police woman-can	man-can
22	*** P the N3	*** with the vacuum	farmer-violin	farmer-fish	farmer-vacuum	farmer-vase	farmer-glue
23	N3	police woman	lady	girl	police woman	dog	pen
24	N3	horse	duck	bird	horse	elephant	dog

Appendix E Full set of derivations

The sequences below show the progress of the candidate set for the variable corresponding to x (the rabbit, in *the rabbit in the big bag*, for example). We use the following codes:

- ✓: a derivation that succeeds
- ✗: a derivation that fails to converge on a referent but does not involve a referential garden path
- ●^{*}: a derivation involving a referential garden path (settling on a referent before failing)

We show all derivations for thresholds $\theta = 2$ and $\theta = 3$, except in those cases where the derivation violates the non-vacuity principle.

E.1 Positive

E.1.1 Positive adjective, THREE-BAGS scene

THREE-BAGS scene, absolute readings

bag(y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✗
bag(y)	+	big _{$\theta=3$} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	● [*]
box(y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✓

THREE-BAGS scene, relative readings

bag(y)	+	rabbit(x) + in(x,y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	uniq(x)	✓
box(y)	+	rabbit(x) + in(x,y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	uniq(x)	✓

E.1.2 Positive adjective, overinformativity scene

OVERINFORMATIVITY scene, absolute readings

bag(y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✓
box(y)	+	big _{$\theta=2$} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✓

OVERINFORMATIVITY scene, relative readings

(None for *box* because both involve violations of the non-vacuity constraint.)

bag(y)	+	rabbit(x) + in(x,y)	+	big _{θ=2} (y)	+	uniq(y)	+	uniq(x)	✓

E.1.3 Positive adjective, both scene

BOTH scene, absolute readings

bag(y)	+	big _{θ=2} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✗
bag(y)	+	big _{θ=3} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✗*
box(y)	+	big _{θ=2} (y)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✓

BOTH scene, relative readings

(None for *box* because both involve violations of the non-vacuity constraint.)

bag(y)	+	rabbit(x) + in(x,y)	+	big _{θ=2} (y)	+	uniq(y)	+	uniq(x)	✓

E.2 Comparatives

E.2.1 Comparative adjective, three-bags scene

THREE-BAGS scene, absolute readings

bag(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✗
box(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x,y)	+	uniq(x)	✓

THREE-BAGS scene, relative readings

bag(y)	+	rabbit(x) + in(x,y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✓
box(y)	+	rabbit(x) + in(x,y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✓

E.2.2 Comparative adjective, overinformativity scene

OVERINFORMATIVITY scene, absolute readings

bag(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x, y)	+	uniq(x)	✓
box(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x, y)	+	uniq(x)	✓

OVERINFORMATIVITY scene, relative readings

bag(y)	+	rabbit(x) + in(x, y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✓
box(y)	+	rabbit(x) + in(x, y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✗

E.2.3 Comparative adjective, both scene

BOTH scene, absolute readings

bag(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x, y)	+	uniq(x)	✗
box(y)	+	cmpr(y, big)	+	uniq(y)	+	rabbit(x) + in(x, y)	+	uniq(x)	✓

BOTH scene, relative readings

bag(y)	+	rabbit(x) + in(x, y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✓
box(y)	+	rabbit(x) + in(x, y)	+	cmpr(y, big)	+	uniq(y)	+	uniq(x)	✗

Appendix F Contrast preference experiment

As stated by the non-vacuity principle, modifiers are most felicitous when there is a non-trivial comparison class; for example, both *the tall man* and *the taller man* are more felicitous in a situation with multiple men (of differing heights) than in a situation involving only one man. Call this the *contrast preference*. But according to the theoretical assumptions we have made, this preference is of a different status for the two sorts of adjectives under consideration. In the case

of positive form gradable adjectives, we assume that readings involving a non-discriminative threshold (one that fails to discriminate among multiple potential referents) are only accessed as a last resort, when no viable interpretation involving a discriminative threshold is available. For positive adjectives, the contrast preference is merely a preference, not a hard constraint. With comparatives, on the other hand, the contrast preference is a hard constraint, encoded in the semantics. In particular, we assume that contexts without a standard of comparison give rise to a presupposition failure. In this appendix, we report on an experiment designed to test these assumptions.

F.1 Methods

F.1.1 Design and materials

Experimental stimuli. There were 24 experimental items. The experimental materials consisted of displays containing four images, one of which, the target object, was highlighted with a red square. All visual displays were accompanied by a written sentence. The sentences were always of the form *This is the big/bigger N*, where N named the type of the target object (e.g., *ladder*). The adjective was either in the positive form (e.g., *big*) or in the comparative form (e.g., *bigger*, see the sample trials in **Figure 17**). There were two independent variables of interest: ADJECTIVE (*big* vs. *bigger*), and CONTEXT (contrast vs. no-contrast). Displays differed according to whether there were one or two objects of the type corresponding to the target object (e.g., one or two ladders, or one or two cups, see **Figure 17**). In NO-CONTRAST displays, there was only one object of the relevant kind (e.g., one ladder). In CONTRAST displays, there were two (e.g., two cups), and the non-target object in the contrast set was smaller. Participants were assigned to one of four

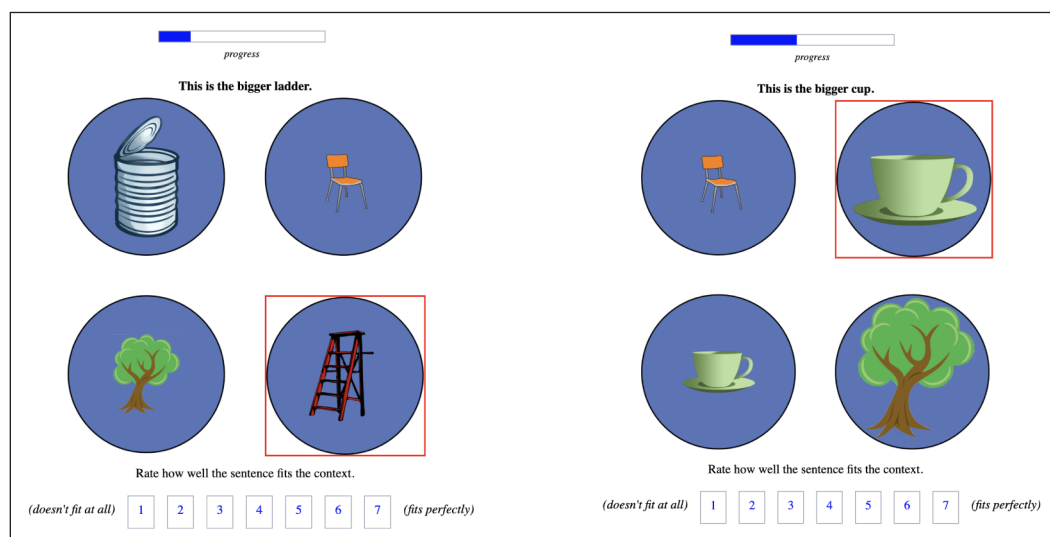


Figure 17: Sample trials for the contrast preference experiment, both involving the adjective *bigger*, one in the no-contrast condition (left), and one in the contrast condition (right).

groups, each corresponding to a different list of stimuli. Each list presented every item in exactly one condition, and the lists were constructed to ensure that every item was shown in every condition, in accordance with a Latin Square design. On each trial, participants were presented with a sentence and a set of four images, one of which was highlighted. The instructions were “Rate how well the sentence fits the context,” and participants were asked to respond on a 1–7 scale, 1 being *doesn’t fit at all* and 7 being *fits perfectly*.

Fillers. There were 24 fillers, each consisting of four images, one of which was highlighted, and a sentence to be judged against the visual context of the images. Thirteen of the fillers were considered attention checks, because the description clearly did or did not fit the highlighted image. For instance, in one case, the description was *This is the green frog*, and while the display contained a brown frog and a green frog, the brown frog was the highlighted referent. In other cases, the description was *These are the glasses* and the display contained only one pair of glasses, which was also highlighted.

F1.2 Participants

Thirty self-reported native speakers of English, recruited through Prolific, took part in the experiment. We excluded from analysis one participant who failed to complete the experiment, and four who answered in an unexpected manner for more than two attention check questions, where “unexpected” was defined as a value less than 4 for a clearly acceptable case or a value of more than 4 for a clearly nonsensical case. Responses from the remaining 25 participants were analyzed.

F1.3 Procedure

At the beginning of the experiment, participants were presented with three practice trials, with differing degrees of acceptability, before moving on to the main experiment.²⁰ Fillers and experimental trials were randomly interspersed, and presented in a different random order for each participant.

²⁰ The practice trials were as follows: The sentence *This is a blue octagon* was paired with a scene with only one blue octagon, which was highlighted. This trial was intended to be fully acceptable; In a second practice trial, the sentence *This is the green shape* was presented, along a scene with two green shapes, one of which was highlighted. This trial was intended to be unacceptable, as the uniqueness presupposition of the definite determiner is not satisfied. Finally, in the third practice trial, the sentence *This is the red circle* was judged against a scene with a red circle and a green circle, and the green circle is highlighted. This practice trial presented the lowest degree of acceptability, since the highlighted object could not be described with either the adjective or the noun in the description.

F.2 Results

The results are shown in **Figure 18**. Clear qualitative differences emerged between the CONTRAST and the NO-CONTRAST conditions, such that judgments were at ceiling in the CONTRAST condition for both positive and comparative adjectives, while acceptability ratings were overall lower in the NO-CONTRAST condition. Importantly, comparative adjectives received lower ratings, compared to positive adjectives in the NO-CONTRAST condition. We constructed a mixed-effects linear regression model with context-fit judgment as the dependent variable and ADJECTIVE, CONTEXT, and their interaction as fixed effects. Items and participants were also included as random intercepts (this was the maximally convergent model). Predictors were coded to sum up to 0 (i.e., 0.5, -0.5). Results displayed a main effect of CONTEXT, such that ratings were higher in the CONTRAST compared to the NO-CONTRAST condition ($\beta = -1.64$, $SE = 0.2$, $p < 0.001$). Crucially, there was also a significant interaction ($\beta = 0.918$, $SE = 0.3054$, $p < 0.01$), such that comparative adjectives were rated significantly lower in the NO-CONTRAST condition ($\beta = -1.06$, $SE = 0.22$, $p < 0.001$), but no significant difference was detected in the Contrast condition ($\beta = -0.05$, $SE = 0.08$, $p > 0.5$).

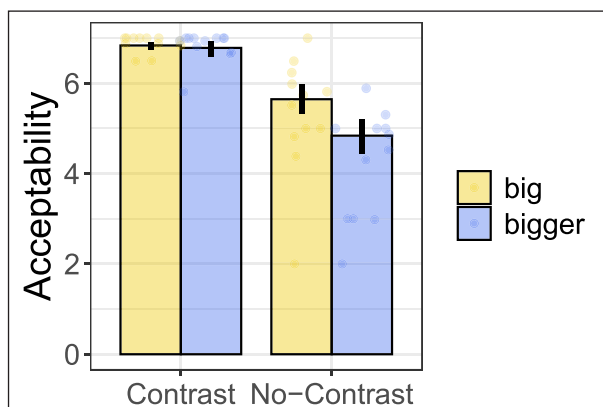


Figure 18: Contrast preference experiment results. Error bars represent 95% bootstrap confidence intervals. Floating dots represent item means.

F.3 Discussion

The results clearly support the conclusion that the preference for the presence of a contrast object in the context is stronger for comparatives than for positives, as shown by the significant ADJECTIVE x CONTEXT interaction. Our theory, in which this requirement is a default interpretive preference for adjectives, but a part of the presuppositional meaning for comparatives, predicts this contrast.

Granted, violations of the contrast requirement for comparatives are not as severe as the effect of mismatching color or object type, as in the nonsensical fillers (which received a mean rating of ≈ 2). This difference may be attributable to the context-sensitive nature of the violation.

Regardless of what else is in the context, a brown frog will never be describable as a green frog. In the case of comparatives in the NO-CONTRAST condition, it is the absence of another object in the scene that causes the violation. We conjecture that other presupposition violations of this kind would give rise to an effect of the same magnitude. Additive particles would be a good testing ground for this conjecture; for example, if participants were asked to assess the statement *This is the man who also has a kitten*, in a context where there is only one man with a kitten, we suspect that the context-fit judgments would be on a par with those that we obtained for comparatives. We leave this to future research; the only point relevant to our purposes here is that the nature of the contrast requirement for comparatives is different, and stronger, than for positive adjectives.

Data accessibility statement

The data and analysis scripts, as well as the auditory and visual stimuli corresponding to the main experiment, the no-modifiers experiment and the contrast preference experiment can be accessed at the following OSF repository: <https://osf.io/asbnj/> (DOI [10.17605/OSF.IO/ASBNJ](https://doi.org/10.17605/OSF.IO/ASBNJ)).

Ethics and consent

The experiments reported in this article are covered under protocol number 4949X submitted to the Institutional Review Board at Boston University. The protocol was approved for IRB exemption on October 17, 2018.

Acknowledgements

We owe an enormous debt of gratitude to the anonymous reviewers of this article and to Florian Schwarz, who slogged heroically through multiple previous drafts, and whose commentary dramatically improved it. We are also grateful to Dylan Bumford and Martin Hackl for helpful discussion, along with the members of the MIT Computational Psycholinguistics Lab, Martin Hackl's Experimental Syntax and Semantics lab at MIT, and audiences at the UCLA Department of Linguistics, the Center for Research in Language at UCSD, the 2023 ESSLLI workshop entitled "Computational and Experimental Explanations in Semantics and Pragmatics", XPRAG 2019, XPRAG-ADJ19, CUNY 32, and the 2019 Stanford SemFest. We gratefully acknowledge support from the National Science Foundation grant BCS-2121074, Elemental Cognition, and the Simons Center for the Social Brain. Last but not least, thank you to Sabrina Tran, Marina Weinstein, and Matthew Briggs for helping us to create the visual stimuli used in the experiments, and to Maria Frey for coding the contrast preference experiment reported in Appendix F.

Competing interests

The authors have no competing interests to declare.

Author contributions

Aparicio and Coppock were in charge of data and stimuli curation, formal (theoretical) analyses, methodology, project administration, as well as the process of writing the original draft. **Aparicio, Levy and Coppock** contributed reviews and editing to the original draft. **Coppock** coded the main experiment and the no-modifiers experiment. **Aparicio and Levy** conducted the statistical analyses. **Aparicio** created the data visualizations. **Levy and Coppock** acquired the necessary funding for this project. The contributor roles are based on the CRediT system taxonomy.

ORCID IDs

Helena Aparicio: <https://orcid.org/0000-0002-8619-1500>

Roger Levy: <https://orcid.org/0000-0002-4493-8864>

Elizabeth Coppock: <https://orcid.org/0000-0001-9987-344X>

References

- Aparicio, H., Chen, C., Levy, R., & Coppock, E. (2021). Granularity in the semantics of comparison. In N. Dreier, C. Kwon, T. Darnell, & J. Starr (Eds.), *Proceedings of Semantics and Linguistic Theory 31* (pp. 550–569). Linguistic Society of America. DOI: <https://doi.org/10.3765/salt.v31i0.5121>
- Aparicio, H., Xiang, M., & Kennedy, C. (2015). Processing gradable adjectives in context: A visual world study. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of Semantics and Linguistic Theory 25* (pp. 413–432). Linguistic Society of America. DOI: <https://doi.org/10.3765/salt.v25i0.3128>
- Beaver, D., & Coppock, E. (2015). Novelty and familiarity for free. In T. Brochhagen, F. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 50–59). Institute for Logic, Language and Computation, University of Amsterdam.
- Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). John Wiley & Sons.
- Bhatt, R. (2002). The raising analysis of relative clauses: Evidence from adjectival modification. *Natural Language Semantics*, 10, 43–90. DOI: <https://doi.org/10.1023/A:1015536226396>
- Bhatt, R. (2006). *Covert modality in non-finite contexts*. Mouton de Gruyter. DOI: <https://doi.org/10.1515/9783110197341>
- Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, 125, 104334. DOI: <https://doi.org/10.1016/j.jml.2022.104334>
- Bumford, D. (2017). Split-scope definites: Relative superlatives and Haddock descriptions. *Linguistics and Philosophy*, 40(6), 549–593. DOI: <https://doi.org/10.1007/s10988-017-9210-2>
- Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, 47, 30–49. DOI: <https://doi.org/10.1006/jmla.2001.2832>

- Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3), 687–696. DOI: <https://doi.org/10.1037/0278-7393.30.3.687>
- Champollion, L., & Sauerland, U. (2010). Move and accommodate: A solution to Haddock's puzzle. In O. Bonami & P. C. Hofherr (Eds.), *Empirical issues in syntax and semantics 8* (pp. 27–52).
- Clifton Jr, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, 2(2), 234–250. DOI: <https://doi.org/10.1111/j.1749-818X.2008.00055.x>
- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to 'overinformative' referring expressions. *Psychological Review*, 127(4), 591–621. DOI: <https://doi.org/10.1037/rev0000186>
- Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6), 725–745. DOI: [https://doi.org/10.1016/0749-596X\(91\)90034-H](https://doi.org/10.1016/0749-596X(91)90034-H)
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178–210. DOI: [https://doi.org/10.1016/0010-0285\(82\)90008-1](https://doi.org/10.1016/0010-0285(82)90008-1)
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58–93. DOI: <https://doi.org/10.1006/jmla.1997.2512>
- Groenendijk, J., & Stokhof, M. (1989). Type-shifting rules and the semantics of interrogatives. In G. Chierchia, B. H. Partee, & R. Turner (Eds.), *Properties, types and meanings. Volume II: Semantic issues* (pp. 21–68, Vol. 2). Springer. DOI: <https://doi.org/10.1007/978-94-009-2723-0>
- Groenendijk, J., & Stokhof, M. (1991). Two theories of dynamic semantics. In J. van Eijck (Ed.), *Logics in AI* (pp. 55–64). Springer. DOI: <https://doi.org/10.1007/BFb0018433>
- Grudzińska, J., & Zawadowski, M. (2019). Inverse linking, possessive weak definites and Haddock descriptions: A unified dependent type account. *Journal of Logic, Language and Information*, 28, 239–260. DOI: <https://doi.org/10.1007/s10849-019-09280-9>
- Haddock, N. J. (1987). Incremental interpretation and Combinatory Categorical Grammar. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 2, 661–663.
- Heim, I. (1982). *The semantics of definite and indefinite noun phrases* [Doctoral dissertation, University of Massachusetts at Amherst].
- Heim, I. (1983). On the projection problem for presuppositions. In D. Flickinger, M. Barlow, & M. Westcoat (Eds.), *Proceedings of the second West Coast Conference on Formal Linguistics* (pp. 114–125). Stanford University Press.
- Heim, I. (1999). *Notes on superlatives* [Manuscript].
- Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57, 129–191. DOI: [https://doi.org/10.1016/0010-0277\(94\)00659-9](https://doi.org/10.1016/0010-0277(94)00659-9)

- Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer Academic Publishers. DOI: <https://doi.org/10.1007/978-94-017-1616-1>
- Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In M. Moroney, C.-R. Little, J. Collard, & D. Burgdorf (Eds.), *Proceedings of Semantics and Linguistic Theory 26* (pp. 836–854). DOI: <https://doi.org/10.3765/salt.v26i0.3937>
- Meier, C. (2003). Embedded definites. In R. van Rooij (Ed.), *Proceedings of the 14th Amsterdam Colloquium* (pp. 163–168).
- Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, 43(8), e12769. DOI: <https://doi.org/10.1111/cogs.12769>
- Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of memory and language*, 110, 104038. DOI: <https://doi.org/10.1016/j.jml.2019.104038>
- Schwarz, F. (2009). *Two types of definites in natural language* [Doctoral dissertation, University of Massachusetts at Amherst].
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2), 109–147. DOI: [https://doi.org/10.1016/S0010-0277\(99\)00025-6](https://doi.org/10.1016/S0010-0277(99)00025-6)
- Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36, 201–216. DOI: <https://doi.org/10.3758/MC.36.1.201>
- Syrett, K., Kennedy, C., & Lidz, J. (2009). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, 27(1), 1–35. DOI: <https://doi.org/10.1093/jos/ffp011>
- Szabolcsi, A. (1986). Comparative superlatives. In N. Fukui, T. Rapoport, & E. Sagey (Eds.), *Papers in theoretical linguistics* (pp. 245–265). MITWPL.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, 268, 1632–1643. DOI: <https://doi.org/10.1126/science.7777863>
- Traxler, M. J., Pickering, M. J., & Clifton Jr., C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39(4), 558–592. DOI: <https://doi.org/10.1006/jmla.1998.2600>
- Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3), 528–553. DOI: <https://doi.org/10.1037/0278-7393.19.3.528>
- van Eijck, J. (1993). The dynamics of description. *Journal of Semantics*, 10(3), 239–267. DOI: <https://doi.org/10.1093/jos/10.3.239>

van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 284–307. DOI: <https://doi.org/10.1016/j.jml.2004.11.003>

van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In *Reading as a perceptual process* (pp. 621–648). Elsevier. DOI: <https://doi.org/10.1016/B978-008043642-5/50029-2>

van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, 45(2), 225–258. DOI: <https://doi.org/10.1006/jmla.2001.2773>