Both Principle B and competition are necessary to explain pronominal disjoint reference effects

Word count: 11398

Breanna Pratley
UMass Amherst
bpratley@umass.edu

Jed Sam Pizarro-Guevara
UMass Amherst
jpguevara@umass.edu

Adina Camelia Bleotu
University of Bucharest
University of Vienna
adina.bleotu@lls.unibuc.ro

Kyle Johnson
UMass Amherst
kbj@umass.edu

Brian Dillon
UMass Amherst
bwdillon@umass.edu

Abstract In two experiments, we investigate the source of the constraint against coreference between a subject and a non-reflexive pronominal object. We contrast two broad claims about what such *pronominal disjoint reference effects* reflect: A *competition* between different means of linguistically encoding a locally coreferent relationship between two nominals in a sentence, and a *constraint* against coreference that is invariable across contexts. In our experiments we studied both the ways in which speakers prefer to express such locally coreferent relationships (Experiment 1), and the interpretations that are compatible with non-reflexive object pronouns (Experiment 2). We find that across different discourse contexts, producers almost categorically avoid producing locally coreferent pronouns (Experiment 1), and generally reject any context that would lead to coreference between a subject and an object in a context inference task (Experiment 2). However, we did see evidence that on a minority of trials, comprehenders allowed a locally coreferent meaning for the pronoun, and assigned it a meaning that is predicted by the competition view. Overall, our results are compatible with the hypothesis that the pronominal disjoint reference effect reflects a semantic constraint against local coreference and binding (a constraint such as Principle B), but that speakers can nonetheless draw inferences about likely meanings for these expressions through general pragmatic reasoning processes.

Keywords: Binding Theory, Anaphora, Experimental Linguistics

1 Introduction

Non-reflexive pronouns are subject to widely-studied constraints on their interpretation. Consider the following example:

(1) a. Dr. Grant$_2$ voted for him$_{1/*2}$
b. Dr. Grant$_2$'s student$_3$ voted for him$_{1/2/*3}$

In (1a), *Dr. Grant* and the pronoun *him* cannot refer to the same individual; but in the apparently minimally different example (1b), they can. This *pronominal disjoint reference*

effect[1] has driven an enormous amount of research into the interface between linguistic form and interpretation. However, the precise theoretical conclusions such pronominal disjoint reference effects license remain obscure.

One view is that they reflect a grammatical constraint that explicitly rules out covaluation between a pronoun and a local referent, as in the original Principle B constraint (Chomsky 1981) or more recent semantically explicit formulations of a similar prohibition (Heim 2007).[2] An alternative view is that the constraint reflected in (1a) is a consequence of a competitive or pragmatic process for determining the optimal linguistic encoding for a given meaning, whereby speakers seek to minimize potential ambiguity (see Levinson 1987, Levinson (1991), Grodzinsky & Reinhart (1993), Kiparsky (2002), and Roelofsen (2010)). The latter hypothesis is attractive, as it dovetails with theoretical proposals about the role that such pragmatic reasoning processes play in language comprehension and production much more broadly (Goodman & Frank 2016; Degen 2023). Insofar as this view invokes very general mechanisms to explain the pronominal disjoint reference effect, it holds the promise of a parsimonious account of pronominal disjoint reference effects. On such a proposal, intuitions about possible interpretations associated with (1a) are ultimately rooted in the fact that there is an unambiguous way of expressing the locally coreferent interpretation (e.g. with a reflexive pronoun), a fact that a canny language user might be able to exploit to draw inferences about the intended or likely meanings for (1a).

In this paper, we present two experiments that seek to contrast these two hypotheses by evaluating the degree to which semantic competition impacts the availability of local covaluation for a direct object pronoun in English. We examine the production and comprehension of pronouns across different discourse contexts to evaluate the degree to which semantic factors that modulate the availability of competing forms correspondingly modulate the possibility of covaluing an object pronoun with a local antecedent, a key prediction of competition-based explanations.

To foreshadow our conclusions, we find evidence for a contextually invariant prohibition against local covaluation for pronouns in English, supporting the hypothesis that speakers encode an explicit constraint that rules out *both* local coreference *and* binding for pronouns (Chomsky 1981; Schlenker 2005; Heim 2007; Jacobson 2007). At the same time, we observe some limited evidence for the kind of pragmatic reasoning over alternative meanings associated with the competition view. In brief, our findings show that the disjoint reference effect in (1a) cannot be entirely reduced to a semantically-oriented competition over alternative formulations of a message, although such competitive reasoning processes do play a minor role in explaining the full pattern of behavior in our experiments.

1.1 Competition and Rule I

Broadly speaking, in a competition-based approach, it is the availability of a 'better' way to convey a given meaning that leads to the pronominal disjoint reference effect in examples like (1a) (Dowty 1980; Grodzinsky & Reinhart 1993; Thornton & Wexler 1999; Kiparsky 2002; Hendriks & Spenader 2006; Reinhart 2006; Van Rij et al. 2010). For example, consider (2). Here, the reflexive pronoun *himself* unambiguously encodes local coreference, since it must be interpreted as referring to the same individual that the local

[1] These effects are commonly called *Principle B effects*, but for present purposes we adopt the term *pronominal disjoint reference effect*, a label that we hope is more theoretically neutral.

[2] See Jacobson (2007) and Schlenker (2005) for similarly semantic versions of Chomsky's account.

subject *Dr. Grant* does. The presence of an effective, unambiguous way of expressing the desired meaning, on this view, somehow 'blocks' comprehenders from assigning a similar meaning to the more ambiguous string *Dr. Grant voted for him*.

(2) Dr. Grant voted for himself. *himself* = Dr. Grant

There are different views on what exactly drives the competition that would result in such a blocking effect. One idea is that it may reflect pragmatic reasoning processes of the sort that underlie a wide range of conversational implicatures (Dowty 1980; Levinson 1987; 1991; Kiparsky 2002), a formalization of the intuition that since a speaker did *not* use a reflexive pronoun, they must have intended some meaning that is not the meaning associated with a reflexive pronoun. A related idea draws on Optimality Theory (Kiparsky 2002; Hendriks & Spenader 2006; Van Rij et al. 2010). Hendriks & Spenader (2006) and Van Rij et al. (2010) propose a 'bidirectional OT' model of the pronominal disjoint reference effect. On this view, listeners take into account the perspective of the speaker in deriving an interpretation of a direct object pronoun, which allows the listener to block the unavailable meaning for a direct object pronoun by recognizing that the speaker *would have* used a reflexive if they had intended the subject and object to corefer. This system posits that simulations of a pragmatically savvy speaker drive listener interpretation. This view makes similar claims about the underlying source of pragmatic enrichment as the rational speech acts model (Goodman & Frank 2016; Degen 2023). Another idea is that the competition reflects a general syntactic or morphological principle that ranks reflexives over pronouns when the former are compatible with the meaning the speaker wishes to convey (Safir 2004; Wyngaerd & Rooryck 2011; Safir 2014).

Grodzinsky & Reinhart 1993's proposal, building on Reinhart (1983; 2006), is central in this competition-based tradition. This proposal starts from the observation that pronouns are capable of getting their value either from the wider discourse context or by being semantically bound by a nominal elsewhere in their sentence (Keenan 1971; Sag 1976; Williams 1977; Gamut 1990; Heim & Kratzer 1998). When a pronoun gets its interpretation from the context, we say that it is a free variable. We say that it is a bound variable when it gets its interpretation from a nominal elsewhere in the sentence. The first important claim of the system in Grodzinsky & Reinhart (1993) is that reflexive pronouns cannot be free variables; they are only interpreted as bound variables. Because bound variables must be c-commanded[3] by the nominal they are bound by, this immediately explains why reflexives can only be interpreted in a way that depends on the value of a c-commanding nominal. That is, there is a contrast in the grammaticality of (3). In the first example in (3a), *himself* refers to the same individual that *Dr. Grant* does, and in the second example of (3a), *himself* has a value that varies with the value given to *everyone*.[4] However, in the ungrammatical examples in (3b), there is no c-commanding nominal that the reflexive can be valued by since *Dr. Grant/everyone* no longer c-commands *himself*.

(3) a. Dr. Grant$_2$ voted for himself$_2$.
Everyone$_2$ voted for himself$_2$.
b. *Dr. Grant$_2$'s mother voted for himself$_2$.
*Everyone$_2$'s mother voted for himself$_2$.

[3] α c-commands β if β is dominated by the phrase that α combines with. It is possible for terms to get a bound variable interpretation without being c-commanded by the term they covary with (Barker 2012), but these possibilities are not available for reflexives.

[4] From this point forward, we represent a pronominal as a bound variable by coindexing it with the nominal it gets its value from.

Non-reflexive pronouns, by contrast, can be both free and bound variables. When free, they get their value from the context of use. This allows pronouns to sometimes have the same referent that another nominal has without being bound to it: Both expressions simply 'point' to the same referent in the context. Thus, there are two ways for *him* to get an interpretation in which it refers to Dr. Grant in (4). In (4a), it is bound by *Dr. Grant*. In (4b), the context gives it the same referent that *Dr. Grant* has, and thus, it said to be *coreferent* with *Dr. Grant*. When a pronominal is interpreted as a free variable, we will give it an index that is not the same as the nominal it has the same referent as. (4) illustrates the two representations that give rise to a reading in which *him* refers to Dr. Grant.

(4) a. Dr. Grant$_2$ said that we should vote for him$_2$. *bound variable*
b. Dr. Grant said that we should vote for him$_2$. 2 = *Dr. Grant; coreference*

Given these distinct mechanisms for interpreting pronouns, to explain the pronominal disjoint reference effect it must be explained why *both* the coreferent *and* bound representations in (5) are unacceptable, since either one could, in principle, deliver a meaning where the pronoun and *Dr. Grant* refer to Dr. Grant.[5]

(5) a. *Dr. Grant$_2$ voted for him$_2$.
b. *Dr. Grant voted for him$_2$. 2 = Dr. Grant

Grodzinsky & Reinhart (1993) argue that this unacceptability arises because of the interplay of two distinct conditions:

(6) A non-reflexive variable cannot be bound by a local argument.

(7) Let S be a clause that contains a pronoun interpreted as a free variable. If the denotation of S remains the same when the pronoun is interpreted as a bound variable, then its interpretation as a free variable is ungrammatical.

The precise definition of "local" in (6) is controversial, and not important for this paper. We will take an argument to be local to a variable if they are both dominated by all of the same clauses. (7), which Grodzinsky & Reinhart call "Rule I," blocks (5b), since the denotation of (5a), where the pronoun is a bound variable, is the same as (5b). (5a), in turn, is blocked by (6), which Grodzinsky & Reinhart call "Principle B."

Taking all the pieces of this system together, we can express the core of this proposal somewhat more succinctly with (8):

(8) **Rule I**
Let S be a clause that contains a pronoun. If the denotation of S remains the same when the pronoun is replaced by a local reflexive, then S is ungrammatical.

By "local reflexive" we mean a reflexive that is local to the argument that binds it, in the same sense that "local" is used in (6). One of the issues that clouds the idea that reflexives and pronouns compete to express the same meaning is that reflexives appear to have more than one way of finding antecedents. Pollard & Sag (1992) argue that only one of those ways is limited by the locality condition relevant for the account of the pronominal disjoint reference effect that we are investigating. The other way likely involves a very different semantic relationship: logophoricity (see Charnavel (2019) for motivation

[5] We reserve the term *covalued* to refer to any situation where a pronoun and an antecedent have the same semantic value.

and discussion). It is only those reflexives whose interpretation is non-logophoric that compete with pronouns.

This is spelled out in (9) and (10).

(9) A local reflexive can only be interpreted as a bound variable.

(10) The binder of a local reflexive must be local to the reflexive.

(10) is known as Principle A. We will call (8) "Rule I." Like the Rule I in Grodzinsky & Reinhart, it prevents a pronoun from being used to express a bound variable meaning when the binder is local. It does so by making the interpretation of a pronoun compete directly with the interpretation of a reflexive in these local contexts. If reflexives are always bound variables, then (8) becomes coextensive with (7).

This revision to Grodzinsky & Reinhart 1993 allows us to dispense with (6) but largely preserve both their empirical results and the analytical structure of their analysis. It also helps highlight the similarities between their proposal and other proposals whereby the availability of a reflexive form blocks the locally covalued interpretation of a direct object pronoun (Kiparsky 2002; Safir 2004; Hendriks & Spenader 2006; Van Rij et al. 2010). Whereas (6) blocks (5a) according to Grodzinsky & Reinhart 1993, (8) does on our version since the denotation of (5a) is the same as the denotation of (11).

(11) Dr. Grant$_2$ voted for himself$_2$.

We call the combination of (9), (10), and (8) the "Competition View." Roughly speaking, it claims that if a locally-bound reflexive can be used to express a meaning, ϕ, then ϕ cannot be expressed with a pronoun in place of the reflexive.[6] Grodzinsky & Reinhart 1993's approach makes clear predictions and has been heavily investigated, and so it provides an ideal starting point for our investigation. As our presentation above underscores, the Competition View reflects a broad, influential idea shared by a range of theories, insofar as it incorporates the claim that a "more specific" form can sometimes block the meaning associated with a more general form (Dowty 1980; Levinson 1987; Grodzinsky & Reinhart 1993; Kiparsky 2002; Hendriks & Spenader 2006; Reinhart 2006; Van Rij et al. 2010).

The Competition View approach has guided much research on first language acquisition. Children learning a range of languages often fail to exhibit pronominal disjoint reference effects, a phenomenon known as the Delay of Principle B effect, or the Pronoun Interpretation Problem (Chien & Wexler 1990; Avrutin & Wexler 1992; McKee 1992; Philip & Coopmans 1996; Baauw 2013). For example, children up to six and a half years of age gave non adult-like responses to examples like (12a) at a high rate, endorsing this as an acceptable description of images where Mamma Bear is touching herself. In contrast, even very young children interpreted examples with reflexives, like in (12b), in an adult-like fashion at very high rates (Chien & Wexler 1987). Results like these have been taken to suggest that constraints on pronouns and constraints on reflexives develop

[6] A subtle distinction between Grodzinsky & Reinhart (1993) and our reframing of their view concerns what the local coreferent interpretation of a pronoun is competing with. In Grodzinsky & Reinhart (1993), the pronoun's interpretation is competing with an interpretation that involves a locally bound variable, whereas on our framing it is competing with the interpretation of a locally bound reflexive. On our view, but not Grodzinsky & Reinhart (1993)'s, the pronoun is competing with actual alternative sentences with a reflexive in them, and is therefore subject to whatever other grammatical constraints govern where reflexives can be. By contrast, Grodzinsky & Reinhart (1993)'s formulation of the constraint has pronouns competing against abstract logical forms. This difference won't matter for what follows.

on different time courses, potentially indicating that they reflect qualitatively different mechanisms in the child's developing grammar.

(12) a. This is Mamma Bear. This is Goldilocks. Is Mamma Bear touching her?
b. This is Mamma Bear. This is Goldilocks. Is Mamma Bear touching herself?

Interestingly, children performed in a more adult-like fashion earlier with quantificational antecedents (e.g. *every bear* instead of *Mamma Bear*). Quantificational antecedents require that the pronoun they covary with be interpreted as a bound variable. The children's differential sensitivity to pronominal disjoint reference effects with coreferential and quantificational antecedents has thus been taken as evidence for the distinction between binding and coreference in the child's developing grammar. The former is syntactically regulated by (6), while the latter is pragmatically regulated by (8). Adding to this general picture, it has been shown that individual variation in Theory of Mind and inhibition in children predicts the rate of adult-like responses for object pronouns (Kuijper et al. 2021).

The overall pattern of results seen in Chien & Wexler 1987's study could be explained by supposing that children acquire purely syntactic constraints on binding at a young age. This would yield adult-like interpretations for reflexive pronouns and pronouns with quantificational antecedents. The selective delay of the pronominal disjoint reference effect with non-quantificational antecedents, then, would be attributed to a delay in the acquisition of the pragmatic competence or working memory resources necessary to implement Rule I or the condition expressed in (8) (Grodzinsky & Reinhart 1993; Thornton & Wexler 1999; Hendriks & Spenader 2006; Reinhart 2006; Kuijper et al. 2021).

In this way, the delay of pronominal disjoint reference effects, and specifically the quantificational asymmetry, has been interpreted as evidence that syntactic and pragmatic or interface competences develop on different time courses, implying that they are dissociable aspects of grammatical knowledge. This, in turn, lends support to a Competition View account of the pronominal disjoint reference effect.

However, this conclusion has not gone unchallenged. Elbourne (2005) argued that classic experimental demonstrations of the quantificational asymmetry effect suffer from an experimental confound. Coreferential antecedents systematically had higher salience in these experiments, which could lead children to select these interpretations at a higher rate. Subsequent experimental research showed that the experimental context does seem to matter quite a bit. Conroy et al. (2009) and Spenader et al. (2009) showed that when the salience of the quantificational and coreferential antecedents were controlled for, children behaved in a way that suggested that they did exhibit pronominal disjoint reference effects in an adult-like fashion in both contexts. Moreover, recent work by Pinto & Zuckerman (2018) showed that children's behavior was more adult-like in a more engaging coloring task than in a picture selection task, a finding which further supports the idea of task context sensitivity.[7]

1.2 Principle B

The underlying intuition behind the Competition View is that there does not need to be any explicitly stated grammatical constraint that rules out coreference between a pro-

[7] Additionally, as pointed out by Sciullo & Agüero-Bautista (2008), while the Delay of Principle B effect seems to show up in languages like English or Dutch (e.g., see Chien & Wexler 1990 for English or Baauw 2013 for Dutch) it seems to be absent or restricted to special structures in languages with clitics (e.g., see Baauw et al. (1997); Baauw (2013) for Spanish, McKee 1992 for Italian, Hamann et al. 1997 for French). This has led researchers to posit a *Clitic Exemption Effect* for Principle B. Interestingly, Klobučar et al. (2025) recently found no evidence for a Delay of Principle B in children even in a language without clitics, such as German. Rather, pronouns in general were found to be challenging. This suggests that an explanation other than the presence of clitics is needed to account for why Delay of Principle B effects emerge only in some languages.

noun and a local antecedent—this observed restriction simply falls out of how speakers and listeners reason over the various alternative ways of expressing a message, and how such reasoning biases interpretation in context. While this is an attractive possibility, not all accounts of the pronominal disjoint reference effect adopt this architecture, especially the claim that the distinction between coreference and binding is central, and that Rule I and the conditions on the interpretation of reflexives in (9) and (10) drive the competition. An alternative view is that English speakers have a single grammatical constraint that directly rules out both local coreference and binding for non-reflexive pronouns (Bassel 2024; Chomsky 1981; Heim 2007; Jacobson 2007; Schlenker 2005). We express this idea in the way that Heim (2007) does. She offers a single-mechanism theory of pronominal disjoint reference effects that simply requires that a pronoun not have the same semantic value as a local c-commanding nominal. Importantly, covaluation is general enough to encompass interpretative relations expressed both by binding and discourse coreference.

In Heim's system, it is necessary to make reference to the idea that quantificational nominals have a syntactic representation in which they bind silent variables, so-called "traces" (May 1977; 1985; Heim & Kratzer 1998). Under this assumption, a sentence like *Every man said that we should vote for him* has the two representations in (13).

(13) a. Every man$_2$ t_2 said that we should vote for him$_2$.
b. Every man$_2$ t_2 said that we should vote for him$_3$.

The silent variable bound by *every man* is represented by *t*. In (13a), *him* is also a variable bound by *every man*, and in (13b) it is a free variable, referring to an individual determined by the context. In (13a), *him* and *t* have the same semantic interpretation; they are "covalued." Compare (13) to (14).

(14) a. Dr. Grant$_2$ t_2 said that we should vote for him$_2$.
b. Dr. Grant$_2$ t_2 said that we should vote for him$_3$.

As in (13a), *him* is covalued with *t* in (14a), and it is interpreted as a bound variable. In (14b), by contrast, *him* is a free variable, and has whatever value the index 3 has. If that index is Dr. Grant, then *him* corefers with *Dr. Grant*. In that scenario, *him* is also covalued with *Dr. Grant*; they both refer to Dr. Grant.

Heim's proposal is that there is a condition that prevents non-reflexive pronouns from being covalued with local, c-commanding, nominals. Like her, we call this Principle B.

(15) Principle B
A pronoun cannot be covalued with a local c-commanding nominal.

This can be seen as a semantically explicit version of Chomsky's Principle B, making reference to covaluation in order to rule out both local coreference and binding. A pronoun is "covalued" with another nominal if the interpretation of that pronoun's referential index has the same semantic value as the nominal; that is, they either corefer or are the same variable. Similar developments can be seen in other more recent approaches (Schlenker 2005; Jacobson 2007; Bruening 2021).

The "Semantic Principle B" view and The Competition View thus encapsulate two different ideas about the source of the pronominal disjoint reference effect. On the former, it is an explicit grammatical constraint on what object pronouns can be covalued with, and on the latter, it results from a competition whereby the presence of a reflexive pronoun "blocks" the missing meaning. We seek to find evidence that helps decide between these views.

1.3 Can pronominal disjoint reference effects be alleviated through discourse?

While the Competition View has been extensively investigated in first language acquisition, the present paper focuses on the predictions it makes for how adult L1 English speakers interpret object pronouns. The Competition View broadly predicts that the strength of a pronominal disjoint reference effect can be weakened or modulated by discourse or semantic factors. In particular, it predicts disjoint reference effects to be weakened in discourse contexts where a reflexive for some reason is disfavored, since that should weaken the reflexive form's ability to block the locally covalued reading for a pronoun.

Evans sentences are a widely considered class of examples thought to have the necessary properties to test this prediction (Evans 1980). Here, we focus on Evans sentences in two specific contexts, exemplified in (16)

(16) a. **Coreferential context**:
Dr. Ellie Sattler said that everyone voted for Dr. Alan Grant, but she lied...
b. **Bound context**:
Dr. Ellie Sattler said that everyone voted for themselves, but she lied ...
... Only Dr. Grant voted for him.

The difference between these contexts is that they distinguish two distinct interpretations of the locally covalued pronoun in the critical sentence, *Only Dr. Grant voted for him*. When *him* in this sentence is a free variable that corefers with *Dr. Grant*, it gets an interpretation that can be paraphrased by (17a). By contrast, when *him* is bound by *Dr. Grant*, it gets an interpretation that (17b) paraphrases.

(17) a. No one but Dr. Grant voted for Dr. Grant.
b. No one but Dr. Grant voted for themselves.

These interpretations arise because of the way that *only* interacts semantically with the focused *Dr. Grant*. Because of the discourse structure of the critical context sentence (*everyone voted for Dr. Alan Grant/ themselves*) and the clause that precedes *but*, the VP of the critical context sentence must have the same denotation as the VP of the complement of *say* (see Rooth 1992 and Schwarzschild 1999 for details). For these reasons, the **Coreferential Context** leads to the coreferent interpretation of *him*, as in (17a), since the critical context sentence encodes reference to Dr. Alan Grant; and the **Bound Context** leads to the bound variable interpretation of *him*, as in (17b), because the critical context sentence encodes a bound variable meaning. In this way, then, these contexts serve to distinguish two distinct meanings that both result in local covaluation between Dr. Grant and the direct object pronoun in the critical sentence.

How do these contexts help arbitrate between competing theories of the disjoint reference effect? On the Competition View, locally covalued interpretations of the pronoun should become (more) available in the Coreferential Context in (16). By contrast, since the critical sentence in the Bound Context expresses a bound variable interpretation of the pronoun, reflexives should be strongly favored, and the locally covalued interpretation of the pronoun should be blocked by Rule I in (8), for example. Note that replacing the pronoun with a reflexive in (16) creates an interpretation that is, by hypothesis, only compatible with the Bound Context.

Rule I therefore predicts that *him* in (16) should lose the competition with *himself* in the Bound Context. In the Coreferential Context, however, the pronoun does not compete with a reflexive, and so Rule I predicts that covaluation should be available. More gener-

ally, if the Coreferential Context weakens a producer's preference for using a reflexive, it should correspondingly weaken the strength of the pronominal disjoint reference effect.

On a Semantic Principle B approach (Heim 2007; Jacobson 2007), local coreference in the Coreferential Context should be ruled out just the same as local binding in the Bound Context. The pronoun is locally covalued in both contexts, and hence is in violation of the constraint against covaluation between a pronoun and a local antecedent in (15). The fact that the preceding context and focus structure of the critical sentence differentiate the contribution of the Bound and Coreferential readings is not predicted to alleviate this effect, since either interpretation ultimately yields illict local covaluation of the pronoun.[8]

For this reason, whether a locally covalued interpretation of the pronoun is available in the Coreferential Context is of significant theoretical importance. If it is available, it would yield support for the Competition View. If it is not available, however, then this would lend support to theories that posit a contextually invariant constraint against local covaluation of pronouns, such as Heim 2007's approach.

Despite their theoretical importance, however, Evans sentences have been relatively understudied experimentally. Intuitions about whether a locally coreferential reading of a pronoun in a Coreferential Context varies among authors. Heim (2007) reports that many judge these to be of "intermediate grammaticality" status; Roelofsen (2010) states that he and many speakers judge this reading to be marginal, departing from Reinhart's original judgments of availability. See also Jacobson 2007.

Although reports of speaker intuition seem to vary, at least some experimental work has yielded some evidence that a locally covalued interpretation of the pronoun is available. Verbuk & Roeper (2010) investigated the interpretation of sentences similar to Evans sentences as in (18):

(18) One morning, Mermaid, Pirate, and Cowboy were going to take a bath. Mermaid had a yellow rubber duck, Pirate had a blue one, and Cowboy a green one. Pirate washed Mermaid. Mermaid didn't wash Cowboy, so Cowboy washed him. Who did Cowboy wash?

The adult controls in their experiment only gave disjoint responses to examples like (18) 25% of the time, while 7-year olds gave disjoint reference responses approximately 50% of the time. This is tentative evidence that Evans sentences may license locally covalued readings of the pronoun, although caution is warranted. The experimental contexts tested by Verbuk & Roeper selectively raise the discourse salience of *Cowboy*, and could contribute to strategic responses on the part of participants (Elbourne 2005).

1.4 The current study

Here we investigate the production and comprehension of Evans sentences. We report the results of two experiments that evaluate whether Coreferential Contexts, as in (16),

[8] Heim (2007) acknowledges that a straightforward application of her theory predicts this effect. However, she takes the availability of a locally covalued interpretation of the pronoun in the Coreferential Context to be acceptable, and proposes a method of accounting for its availability. She suggests that the term associated with *only* or *even* is syntactically embedded inside a phrase that is headed by "F," the formative that gives this term its focused meaning. As a consequence, she suggests that the focused term doesn't c-command anything, like the following object pronoun, that is not within FP. On this view, *Dr. Grant* doesn't c-command *him* in (16); as a consequence, her Principle B doesn't apply to them. We view this additional elaboration of her core proposal to be *ad hoc*, and since our results show that the predictions of the Competition View for examples like (16) are not met, unnecessary for the present analysis.

license locally covalued pronouns, as predicted by the Competition View. We investigate this question in production, with a cloze production task in Experiment 1, and in comprehension, with a 'Guess the Context' task in Experiment 2. In both experiments, we test whether locally covalued pronouns are associated with Coreferential contexts.

We explicitly chose *not* to ask for acceptability judgments, because informally reported intuitions have been variable and unclear in the preceding literature (Schlenker 2005; Heim 2007; Jacobson 2007; Roelofsen 2010), suggesting that judgment data *per se* are unlikely to provide unambiguous evidence about the acceptability of locally covalued readings of the pronoun for English speakers.

Instead, we ask two key questions in our experiments. First, do people ever *produce* pronouns when a locally coreferent meaning is intended? Secondly, do people associate a locally covalued pronoun with Coreferential Contexts?

A Competition-based theory predicts that 1) speakers should produce locally covalued pronouns in contexts where a reflexive is less strongly associated with the target meaning, as in the Coreferential Context, and 2) speakers should associate a locally coindexed pronoun with a Coferential Context over a Bound Context. However, a theory rooted in a Semantic Principle B predicts that 1) speakers should avoid producing locally coindexed pronouns in Bound and Coreferential contexts alike and 2) speakers should not associate locally coindexed pronouns with either context.

2 Experiment 1: Cloze production experiment

Our goal in Experiment 1 was to better understand what referring expressions English speakers use when producing Evans sentences in different discourse contexts. In this experiment, participants were given a modified Cloze production task. Participants were shown a series of displays featuring simulated text messages (Kroll 2020), and asked to complete the natural exchange from the perspective of the 'texter' by filling in the blank. See Figure 1 for an example of a critical trial in experiment 1.

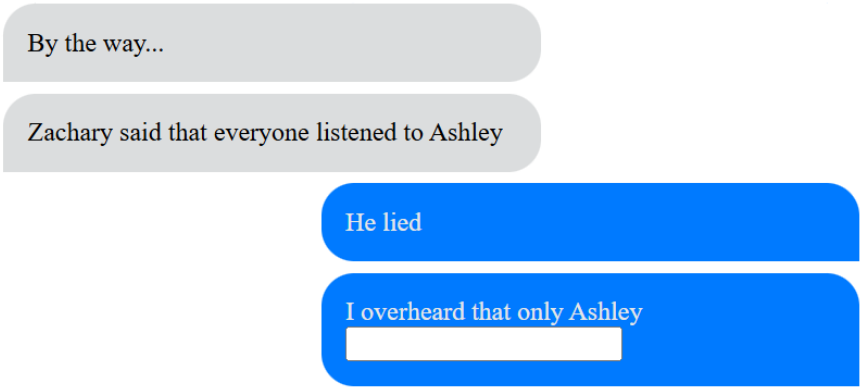

Figure 1: An example of a critical trial in Experiment 1

The critical target text message was embedded in either a Coreferential or Bound context (second gray text bubble), but with a portion of the critical target sentence left blank (blue text bubble with a text field). We expected participants in this task to fill in the blank with reference to a critical antecedent phrase in the context, given the focus structure of the overt portion of the target sentence.

We compared the participants' behavior for three different types of contexts. In the **Bound** condition, the context expressed a local binding relationship between the subject and the object with a reflexive pronoun. In the **Coreferential** condition, the antecedent phrase expressed reference to a target character with a proper name. These are the two contexts provided in (16). In addition to these two contexts, we created a third condition, the **Non-Coreferential** condition. This was identical to the Coreferential context, with the sole change that the name in the target sentence was not mentioned anywhere before.

We expected participants to produce primarily continuations with some type of local covaluation relationship between the subject and the object of the target sentence in the Coreferential and Bound conditions, but not in the Non-Coreferential condition, which served as a control.

Our target of investigation was how the choice of referring expression used to express a local covaluation relationship would differ across conditions. Semantic Principle B predicts that comprehenders should not use pronouns to express this relationship in either Coreferential or Bound condition, since local covaluation of a pronoun with its subject is prohibited in both cases. In contrast, the Competition View predicts that the probability of using a pronoun to express local covaluation should be higher in the Coreferential condition than in the Bound condition. This is because in the Coreferential condition, reflexive pronouns are expected to be less available to express the target meaning than in the Bound condition.

2.1 Participants

We recruited 36 American participants with English as their L1 through Prolific, a crowdsourcing platform (https://www.prolific.com/). Each participant received 6 USD as compensation.

2.2 Materials

We implemented six conditions in a 3×2 crossed factorial design with two factors: Context and Prompt Type. The Context factor had three levels: **Bound**, **Coreferential**, and **Non-Coreferential**. The Prompt Type factor manipulated the size of the blank that participants were supposed to fill out. Participants were either given a prompt that contained the critical repeated verb (+**Verb**) or did not (**–Verb**). In Table 1, we provide an example stimuli by condition.

Condition	Context	Fill-in-the-blank Target
Bound: *x* listened to *x*	By the way, Zachary said that everyone listened to themselves.	He lied! I overheard that only Ashley {_____./listened to ____.}
Coreferential: *x* listened to Ashley	By the way, Zachary said that everyone listened to Ashley.	He lied! I overheard that only Ashley {_____./listened to ____.}
Non-Coreferential: *x* listened to Jacob	By the way, Zachary said that everyone listened to Jacob.	He lied! I overheard that only Ashley {_____./listened to ____.}

Table 1: Example stimuli by condition for Experiment 1.

We created 48 critical items in this design. These items were evenly distributed via Latin square and were randomized along with 48 filler items. The structure of all items followed the same pattern. There were two text messages from another speaker that set the context. The context always introduced a statement made by a third character (e.g., *Zachary said that everyone listened to themselves*.). The target text messages immediately followed. The first message was a denial or affirmation of the truth of this character's statement (e.g., *He lied*!). The second message always contained a focus particle attached to the subject in order to create a focus contrast with an antecedent phrase in the context. The polarity of the focus particle was counterbalanced across items; for half of the items, the target texts denied the antecedent and used *only* as the focus particle, and for the other half, they affirmed the antecedent and used *even* as the focus particle.

The 48 fillers were balanced for whether the first message made use of the quantifier *most* or the quantifier *everyone* in the first message (e.g., *Zachary heard that most people/everyone loved the new film*), whether the response message contained an embedded sentence starting with a focus particle *only* or *even* (*I overheard that only/even Brianna...*), and whether it made use of a single proper name or two coordinated proper names (*I overheard that only Michael and Alexis are...*), as well as for whether the second message could be continued through a nominal element or a verbal and nominal element (similar to the practice items in Figures 2 and 3).

2.3 Procedure

Participants were asked to pretend that they were at a work party, texting their closest co-worker about their other co-workers. They were told that there were around 24 people at the party and that no two people at the party had the same name. This was done to ensure coreferential interpretations of repeated names.

Participants were given two main tasks: (i) to read the text messages (their friend's messages are boxed in gray, while their own messages are in blue), and (ii) to write a response to their friend's messages by typing a continuation in a text box.

To familiarize participants with the task, we introduced them to a practice item where they had to fill in a nominal element (possibly a pronoun/proper name), such as the one shown in Figure 2, as well as a practice item where they had to fill in a verb and its nominal argument (possibly a pronoun/proper name), such as the one in Figure 3. After the practice trials, participants engaged with the actual experiment.

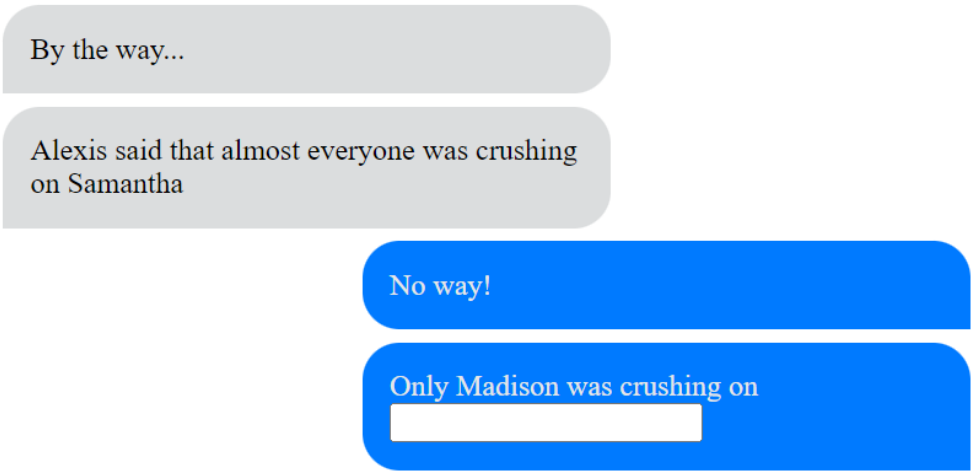

Figure 2: Practice item requiring nominal continuation in Experiment 1

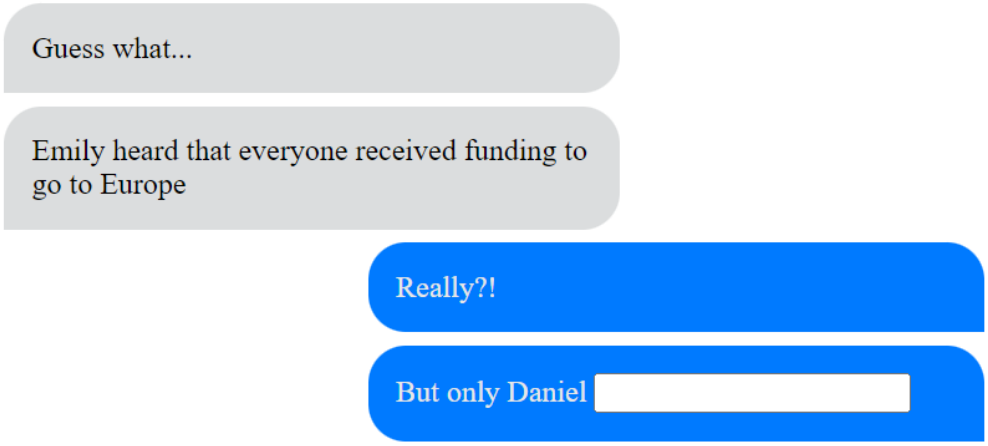

Figure 3: Practice item requiring verbal and nominal continuation in Experiment 1

2.4 Analysis

Participants' completions of the target sentence were recorded. To analyze the completions, participants' responses were hand-coded along a number of dimensions. First, each participant's response was categorized as *On Target* or not. If a continuation was judged to be parallel to the antecedent phrase in the context (e.g., *everyone listened to themselves/Ashley/Jacob* in Table 1), then it was considered On Target. Otherwise, it was *Not On Target*. An example of both types of completions are given in (19):

(19) **Context:** Zachary said that everyone listened to themselves.
a. **On Target production:** He lied! Only Ashley listened to herself.

b. **Not On Target production:** He lied! Only Ashley listened to what Zach said.

For On Target productions, responses were categorized according to the form of the referring expression used: reflexive pronouns, pronouns, proper names, epithets (e.g., *that stupid boy*, *his dumb face*), implicit objects, and other. On Target 'other' productions occurred when a participant offered a continuation that was judged discourse parallel in the relevant sense, but which had significantly different surface syntax (e.g., '*Jacob said that most people photographed themselves*' → '*He lied! Only Madison takes selfies*').

To analyze our data, we fitted a series of Bayesian hierarchical logistic regression models in R (R Core Team 2020) using `brms` (Bürkner 2017), a front end of the Stan language for Bayesian estimation of model parameters (Gelman et al. 2015). To extract estimated marginal mean response proportions for our conditions, we used `emmeans` (Lenth 2025). Each model contained fixed effects for Context and Prompt Type. The contrast coding used in our analysis is summarized in Table 2.

Experimental condition	Prompt Type	Local covaluation	Binding
Bound, +Verb	-1	-1	+1
Coreferential, +Verb	-1	-1	-1
Non-Coreferential, +Verb	-1	2	0
Bound, -Verb	1	-1	+1
Coreferential, -Verb	1	-1	-1
Non-Coreferential, -Verb	1	2	0

Table 2: Contrast coding used for all regression models in Experiment 1.

We used Normal (0,1.5) as the priors for all the fixed effects, and Normal (0,10) as the prior on the the intercept. These are uninformative priors as they do not place strong constraints on the model's predictions, and incorporate very little knowledge about what makes a plausible distribution of the dependent measures. We used LKJ(2) as the prior for our correlation matrix. For each model, we ran four Monte Carlo Markov chains in parallel, with 8,000 samples each. The first 4,000 were always discarded as part of warm-up. For all the models reported below, the R-hat statistic was at 1.0. No divergences were observed.

We fit three different models with these parameters. First, we analyzed the effect of our experimental manipulations on the probability of an On Target response. Next, we analyzed the effect of our experimental manipulations on the probability of a Pronoun response, and the probability of a Proper Name response. For these latter two models, we limited our analysis only to On Target responses. We used the `emmeans` package to extract predicted values by condition (Lenth 2025).

2.5 Results

2.5.1 Descriptive summary

Figure 4 presents a descriptive summary of the On Target results by context. We present only continuations in three categories: Reflexives, Pronouns, and Proper Names, as there were few to no responses in any other response category.

In the **Bound** condition, we saw a near categorical preference for reflexive completions, as expected. In the **Non-Coreferential** control condition, participants produced largely Proper Name and Pronoun continuations, with a preference for the former. They

produced no On Target Reflexive continuations, as expected, given that this context did not cue a continuation where the subject and object were covalued. Across contexts, we failed to see any consistent influence of the presence of a verb in the prompt on the form of the continuation.

In the critical **Coreferential** condition, unexpectedly, Reflexives made up the majority form of On Target responses. Despite this preference, our contextual manipulation was nonetheless successful at weakening the preference for a Reflexive continuation, as predicted by the Competition View. We also saw a significant number of Proper Name continuations.

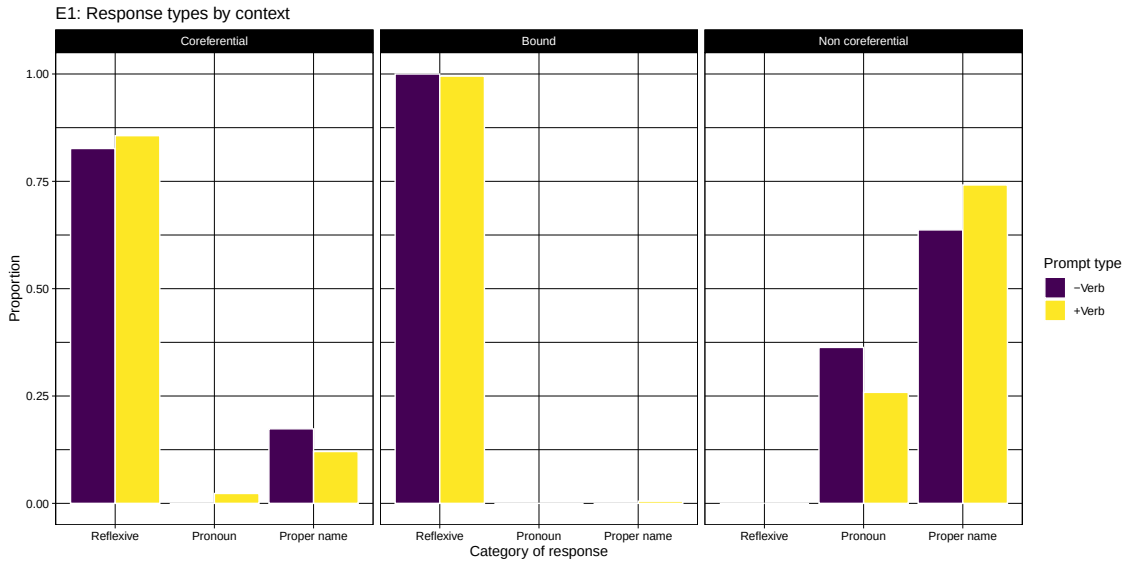

Figure 4: Analysis of On Target responses for Experiment 1, showing the rate of Reflexive, Pronoun, and Proper Name continuations by condition.

2.5.2 On Target response analysis

Figure 5 presents the overall observed mean rate of On Target responses, along with fitted On Target response probabilities and 95% HDI derived from the Bayesian multilevel model. The posterior estimate for all fixed effects, along with their associated 95% highest posterior density intervals are presented in Table 3. If the posterior density interval overlaps with 0, we cannot confidently rule out that it is a null effect (e.g., that 0 is an appropriate parameter setting given our data), and hence do not treat it as meaningful (Kruschke 2014; Schad et al. 2021).

There were more On Target responses when the prompt contained a verb than when it did not. The highest rate of On Target responses was seen in the Non-Coreferential condition; both conditions with local covaluation had fewer On Target responses overall. However, within these two contexts there were significantly fewer On Target responses in the Coreferential context compared to the Bound context. The model revealed that these general trends were qualified by an interaction effect: the tendency for local covaluation contexts to give fewer On Target responses was greater when the prompt did not include a verb.

Effect	$\hat{\beta}$	95% HDI
Local covaluation	0.65	[0.45,0.88]
Binding	0.55	[0.35,0.74]
Prompt	-0.40	[-0.64,-0.19]
Local cov.×Prompt	-0.20	[-0.40,-0.02]
Binding×Prompt	0.16	[-0.04,0.35]

Table 3: Posterior estimates for all fixed effects in the Bayesian multilevel logistic regression model of the probability of an On Target response.

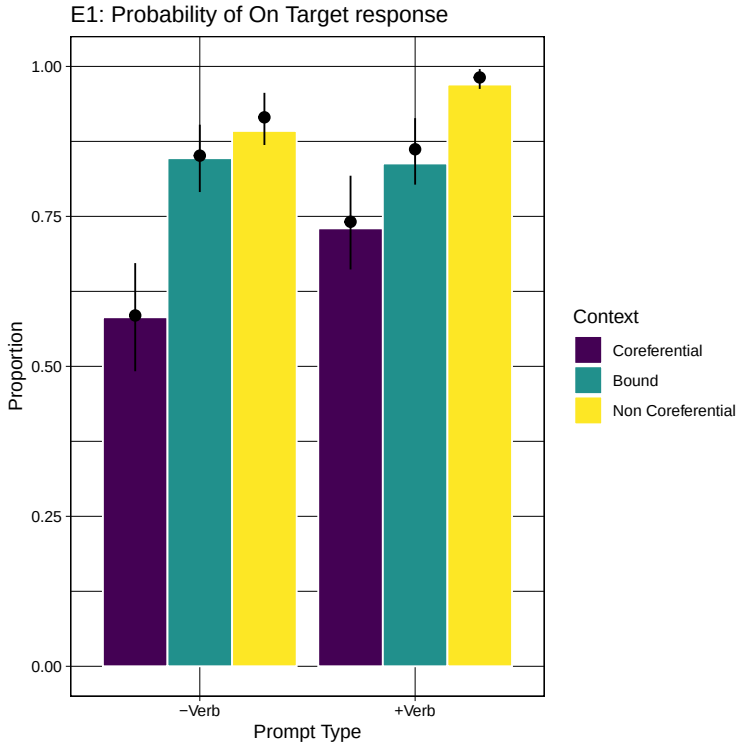

Figure 5: Probability of an On-Target response across conditions. Bars indicate the mean empirical proportion of On-Target responses. Points and error bars indicate the estimated marginal mean for each condition along with its 95%HDI from the fitted Bayesian multilevel logistic regression model.

2.5.3 Pronoun response analysis

A critical part of our analysis concerns the rate of pronoun responses in the Coreferential condition. Figure 6 presents the overall observed mean rate of Pronoun responses within the On Target responses, along with fitted On Target estimated marginal means for each condition along with its 95% HDI derived from the Bayesian multilevel model. The posterior estimates for all fixed effects, along with their associated 95% highest posterior density intervals are presented in Table 4.

There were significantly more pronoun responses in the Non-coreferential conditions than in the other conditions, reflected in the large effect of **Local covaluation** in the regression model. The 95% HDI for all other model coefficients overlapped with 0.

Effect	$\hat{\beta}$	95% HDI	
Local covaluation	3.07	[1.98,4.63]	Pr(β<0)=0
Binding	-0.85	[-3.10, 1.38]	Pr(β<0)=0.78
Prompt	-0.87	[-2.65,0.75]	Pr(β<0)=0.85
Local cov.×Prompt	0.51	[-0.31,1.42]	Pr(β<0)=0.11
Binding×Prompt	2.38	[-0.72,6.8]	Pr(β<0)=0.08

Table 4: Posterior estimates for all fixed effects in the Bayesian multilevel logistic regression model of the probability of a Pronoun response.

Given the fitted model, the estimated marginal mean of a Pronoun response was 3.7×10^{-6} (95% HPD [0,0.001]) in the +Verb prompt, and 4.33×10^{-3} (95% HPD [0,0.02]) in the -Verb prompt condition. The fitted estimates assign a probability of near zero for pronoun responses in the Coreferential context in our experiment. Overall, we observed 4 On Target Pronoun responses in the Coreferential conditions, out of a total of 492 total recorded responses.

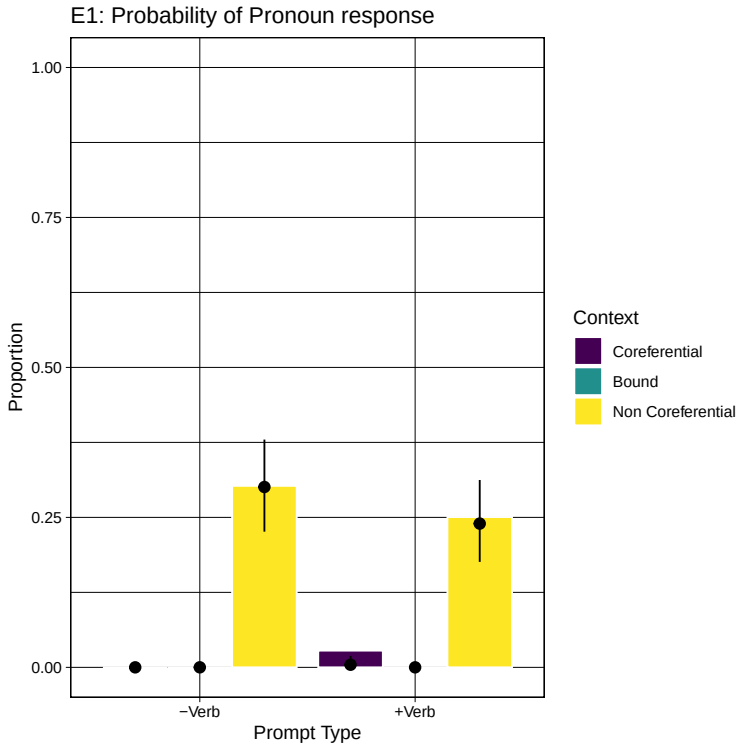

Figure 6: Probability of a Pronoun response across conditions. Bars indicate the mean empirical proportion of Pronoun responses. Points and error bars indicate the posterior mean and 95% HDI by condition from the fitted Bayesian multilevel logistic regression model.

2.5.4 Proper Name response analysis

Figure 7 presents the overall observed mean rate of Proper Name responses within the On Target responses. The accompanying regression model is summarized in Table 5.

Effect	$\hat{\beta}$	95% HDI	
Local covaluation	1.75	[1.41,2.22]	Pr(β<0)=0
Binding	-2.40	[-3.72, -1.45]	Pr(β<0)=1
Prompt	-0.46	[-1.37,0.24]	Pr(β<0)=.89
Local cov.×Prompt	0.02	[-0.34,0.58]	Pr(β<0)=0.50
Binding×Prompt	-.59	[-1.97,0.46]	Pr(β<0)=0.85

Table 5: Posterior estimates for all fixed effects in the Bayesian multilevel logistic regression model of the probability of a Proper Name response.

There were significantly more Proper Name responses in the Non-Coreferential conditions than in the locally covalued conditions. However, within the locally covalued conditions there were significantly more Proper Name responses for the Coreferential condition over the Bound condition.

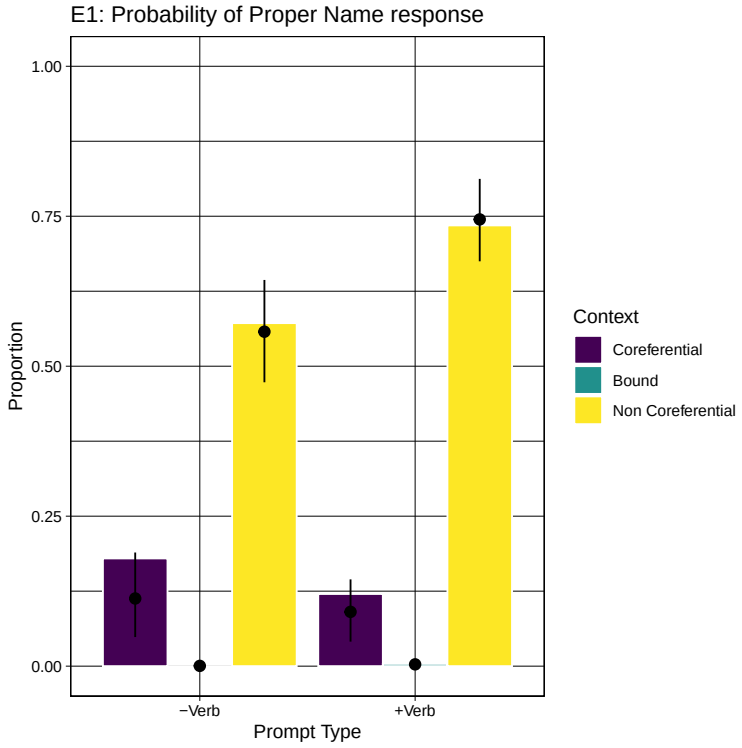

Figure 7: Probability of a Proper Name response across conditions. Bars indicate the mean empirical proportion of Proper Name responses. Points and error bars indicate the posterior mean and 95% HDI by condition from the fitted Bayesian multilevel logistic regression model.

2.6 Discussion

The primary goal of Experiment 1 was to determine what types of referring expressions English speakers use to express a local covaluation relationship between a subject and an object. We tested this in two distinct discourse contexts, a Coreferential context and a

Bound context. We further compared speakers' behavior to a minimally different context where local covaluation was not expected.

In the critical Coreferential context, we observed less than 0.1% Pronoun responses, with the vast majority of participants (92%) not producing a single pronoun response in this condition; in contrast, pronouns were readily produced in our experiment in the Non-Coreferential control context. We found no evidence that the rate of pronoun production differs between Bound and Coreferential contexts, as predicted by the Competition View. The extremely low rate of Pronoun responses in both contexts where the discourse promoted local covaluation between the subject and the object is overall more consistent with a Semantic Principle B view, as the discourse did not seem to alleviate speakers' reluctance to use a non-reflexive pronoun in either context.

We did, however, see two unexpected results in the data. First, an analysis of the rate of On Target responses revealed a sharp drop in the rate of On Target responses in the Coreferential Context. We stress that this result was not predicted *a priori*. Nonetheless, it is potentially informative with respect to the central question of this paper, as it suggests that participants were reluctant to produce *any* locally covalued continuation of this discourse. If this is the right interpretation of this pattern, then it suggests that the prohibition against using a pronoun to express local covaluation between a subject and an object in the critical sentence may be quite strong indeed, insofar as speakers seem to rank creating a potentially incoherent discourse above producing a pronoun. We take up this finding in more detail in the General Discussion.

In addition, we did see an unexpectedly high rate of reflexive continuations in the Coreferential context. On the assumption that reflexives obligatorily express a bound variable relationship (see 1.1 above), this finding is puzzling. This observation is consistent with other recent claims that English speakers can use reflexives to express a locally covalued free variable (Bruening 2021; McKillen 2016). We take up the broader implications of this finding in the General Discussion. However, we note that despite this unexpected finding, we did see that speakers produced far fewer reflexive continuations in the Coreferential context than in the Bound context.

Overall, this result challenges the Competition View, as a decrease in the rate of Reflexive responses in the Coreferential context did not increase the rate of Pronoun responses. Instead, it decreased the rate of On Target responses, and increased the rate of Proper Name responses (which constitute a Principle C violation or Repeated Name Penalty in their own right, as they involve repeating a noun phrase in rapid succession, e.g. *Dr. Grant voted for Dr. Grant*). This overall data pattern is more consistent with a Semantic Principle B view, which suggests that speakers have a strong constraint against any local covaluation between a subject and a non-reflexive pronoun, regardless of context.

3 Experiment 2: Context selection experiment

In Experiment 1, we found no evidence that producers consistently produce locally covalued pronouns in the **Coreferential context** configurations; instead, they overwhelmingly preferred reflexives or proper name continuations. This is, on the face of it, inconsistent with the predictions of the Competition theory, as described above. However, the strength of this conclusion could be challenged. It is possible that participants do grammatically permit locally covalued pronouns in this configuration, but due to some other factor, they did not produce them in this task. For example, it could be that in our stimuli, the target referent was not sufficiently salient to support the use of a non-reflexive pronoun.

Therefore, in Experiment 2, we provide a complementary test of the predictions of the Competition and Principle B views. In Experiment 2, we implemented a "Guess the Context" task. In this task, participants were shown a text message exchange, akin to the ones in Experiment 1. However, rather than asking them to continue the target text message, in this experiment, participants were provided with the target text message and were asked to choose which context was consistent with the message (see example 8 below). Participants were given two discourse contexts as response options; each context encouraged a locally covalued interpretation of the pronoun in the target text message. One of the two contexts was consistent with a bound reading of the pronoun in the target text message, and the other was consistent with a coreferential reading. They could choose one, the other, or both as appropriate contexts for the text message exchange. They were also given the option to reject either context as appropriate for the pronoun. See Figure 8 for an example of a critical trial in experiment 2.

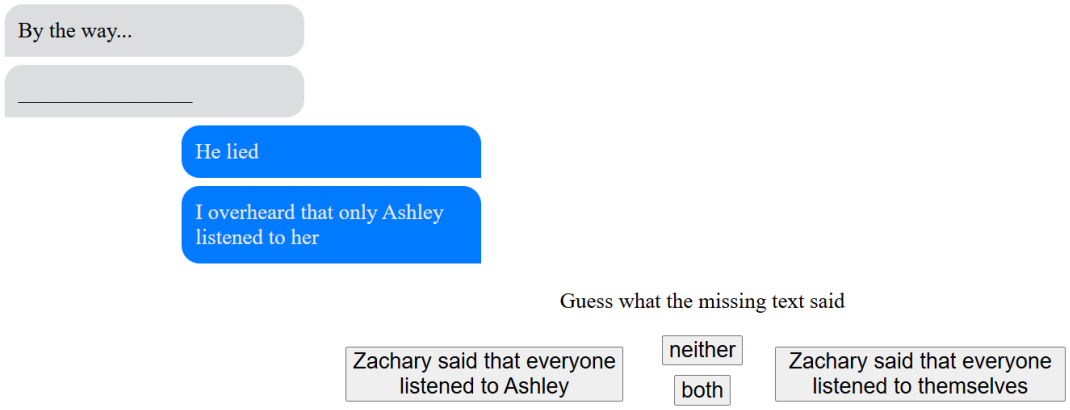

Figure 8: Example of a test item in the Pronoun condition in Experiment 2

The predictions for this task mirror those of Experiment 1. The Competition view leads us to predict that participants will select the coreferential context over the bound context as the missing context in the text exchange when presented with a locally covalued pronoun. In contrast, the Principle B view leads us to predict that participants will reject both contexts as suitable candidates for the missing text exchange.

3.1 Participants

We recruited 55 participants using Prolific and were compensated at a rate of 12 USD per hour. On average, participants took about 25-30 minutes to finish the experiment. Prior to any analyses, we excluded one participant due to a fast completion time (i.e., about 4 minutes).

3.2 Materials

We implemented three conditions in a 3×1 design with a single factor: Sentence Type. Sentence Type manipulated whether the final sentence in the text exchange contained a reflexive, a pronoun, or a proper name in direct object position, thus giving rise to 3 levels: **Reflexive**, **Pronoun**, **Proper Name**.

For each target sentence, participants were given the option of selecting one of four response options for the missing text in the text exchange, as shown in Table 6. One

option had a reflexive in the direct object position; another option had the name used in the critical test sentence in direct object position. Additionally, participants had the option to select *both* or *neither*.

In addition to these three critical conditions, we implemented two parallel control comparisons with a similar design. The goal of these control comparisons was to ensure that participants were behaving as expected in this task. The first control comparison we termed the **Form controls**. Like the critical items, in this set of controls we manipulated the form of the referring expression in the critical prompt sentence. In these trials, both main response options had a name that matched the gender of the referring expression in the critical sentence; One of the response options contained a name which was also in the subject position of the critical sentence, and the other response option contained a different name. This set of controls allowed us to evaluate whether participants performed as expected with all three referring expressions, choosing the context that was consistent with local covaluation of the target referring expression for reflexive and proper name prompts, and choosing the context that was consistent with local disjoint reference for the pronoun prompts. In addition, these controls measured the extent to which a participant would choose a response option containing the same name included in the critical prompt sentence—in other words, this set of controls aimed to ensure that any results in the critical conditions did not simply reflect a shallow name-matching strategy. No inferential statistics were run on these controls.

The second set of controls consisted of **Neither controls**, which sought to determine the extent to which participants actually were sensitive to the overall discourse structure, and to what extent they were willing to choose 'neither' as a response option. Like the Form Controls, both response options contained a name in the direct object position. However, the gender associated with both of these names mismatched the gender of the name or pronoun contained in the critical test sentence. This gender mismatch ensured that both response options would create a non-parallel discourse, and hence, we expected high rates of 'neither' responses in these controls if participants were sensitive to the overall discourse structure of the text exchanges we created. Again, no inferential statistics were run on these controls.

We created 45 item sets spanning these 9 distinct configurations: the 3 critical conditions and the 6 control conditions. The items were evenly distributed via Latin square and were randomized along with 60 filler items, resulting in 5 observations per (critical) condition per participant. See Table 6 for sample items.

The linguistic structure of all items in the experiment followed the same pattern. We again followed the common text convention with grey bubbles indicating messages from an interlocutor, and blue bubbles indicating messages issued by the producer themselves. In the displayed text screens, the interlocutor gave an introductory message (often an interjection), followed by a missing text message. The speaker/texter's response immediately followed, consisting first of a denial or affirmation of the truth of the missing statement (e.g., *He lied*). This was followed by the critical sentence. Each sentence began with a subordinating clause such as *I overheard that*, followed by an embedded clause where the focus particle *only* modified a critical name, followed by a verb. The form of the nominal in the object position of this verb was manipulated such that it was either a proper name, a pronoun, or a reflexive. As in Experiment 1, the polarity of the focus particle was counterbalanced across items: Half of the items contained *only*; half contained *even*.

Critical items		
Context-setting	By the way, Zachary said that everyone listened to	
CONTEXT	**Bound**	themselves.
	Coreferential	Ashley.
Final text	He lied! I overheard that only Ashley listened to	
	Reflexive	herself.
SENTENCE TYPE	**Pronoun**	her.
	Proper Name	Ashley.
Form Control		
Context-setting	By the way, Olivia said that everyone listened to	
CONTEXT	**Prompt subject mentioned**	Christopher.
	Prompt subject not mentioned	Jacob.
Final text	She was being honest! I overheard that only Christopher listened to	
SENTENCE TYPE	**Reflexive**	himself.
	Pronoun	him.
	Proper Name	Christopher.
Neither Control		
Context-setting	Guess what??? Zachary said that everyone voted for	
CONTEXT	**Answer 1**	Daniel.
	Answer 2	Christopher.
Final text	He was being honest. I heard that even Alexis voted for	
SENTENCE TYPE	**Reflexive**	herself.
	Pronoun	her.
	Proper Noun	Alexis.

Table 6: Example stimuli by condition for Experiment 2. Participants could choose between the Bound Context, the Coreferential Context, Both, and Neither.

3.3 Procedure

As in Experiment 1, participants were asked to pretend that they were in a work party setting where multiple people were being discussed, and that there were no two people with the same name. Participants were then told that they and their coworker were texting and gossiping about all the other coworkers. However, their phones were having some technical issues, and as a result, some of the messages from their coworker were coming through blank. Against this backdrop, the participants' task was to guess which text message they thought was the missing one in the conversation.

To familiarize participants with the task, they were given a practice item wherein one of the options was clearly contextually relevant and one was not (see Figure 9). After the practice trial, participants engaged with the actual experiment, where on each trial they were to respond with one of the four offered response options.

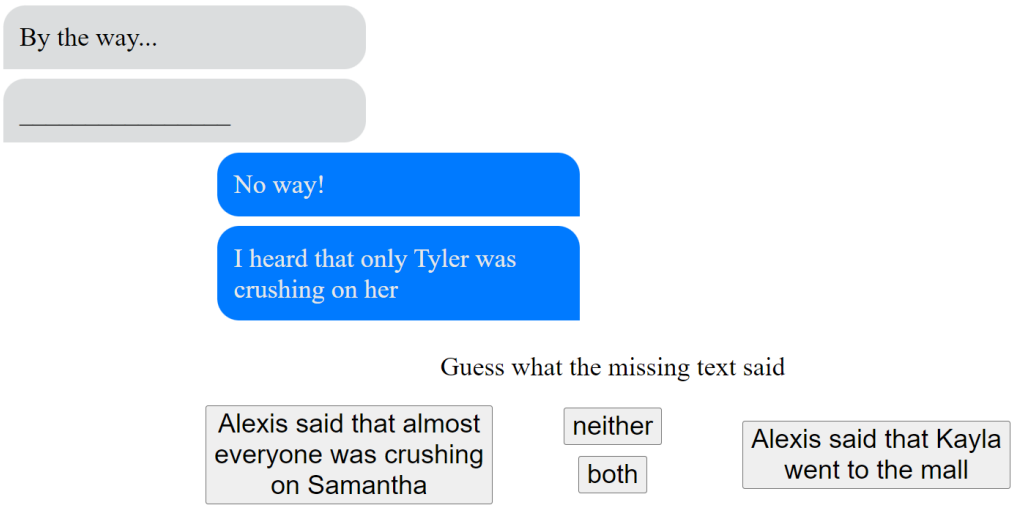

Figure 9: Example of a practice item in Experiment 2

3.4 Predictions

In our critical conditions, both response options offered a context where subject and object nominals in the critical test sentence would be construed as locally covalued. The questions our experiment sought to answer were the following i) will any context be accepted as possible, and ii) will preferences for inferred context vary by the form of the nominal in the critical test sentence?

The prediction of a Semantic Principle B account is simple: there is no way to grammatically license local covaluation of a pronoun and a local subject, therefore, no response option should be appropriate, and 'neither' should be chosen at very high rates (comparable to 'neither' controls) for the pronoun condition.

On the other hand, the Competition view holds that a locally covalued pronoun is a possible interpretation, provided that this interpretation occurs in a context that allows the pronoun to receive a coreferential reading that is different from a bound variable reading. Thus, the prediction of the Competition view is that there should not generally be a high rate of 'neither' responses, but that instead, there should be a high rate of Coreferential responses, and a low rate of Bound responses. By hypothesis, when the target sentence contains a reflexive pronoun, there should be a high rate of Bound responses, a lower rate of Coreferential responses, and a low rate of 'neither' responses.

3.5 Analysis

As described above, inferential statistics were only run on the three critical conditions. We fitted Bayesian hierarchical logistic regression models in R (R Core Team 2020) using `brms` (Bürkner 2017). We fitted three models models aimed at evaluating the predictions laid out above. First, we modeled the probability of a 'neither' response by prompt form (the Neither Model). Second, we fitted two further models to estimate the extent to which participants allowed a coreferential context, given a particular form (the Coreferential Model); and another one to estimate the extent to which they allowed a bound context, given a particular form (the Bound Model).

For the Coreferential and Bound Models, we binarized participants' response options, and used the binarized response as the dependent measure. In the Coreferential Model,

we coded their response as 1 if they chose "both" or the context option that mentioned the name; else, 0. In the Bound Model, we coded their response as 1 if they chose "both" or the context option that contained a reflexive. Otherwise, it was 0.

In both models, we used as fixed effects the referential form of the object nominal in the final text message, dummy coded with **Pronoun** as the reference level. The random effects structure included random intercepts for participants and items, random slopes for the referential form for both participants and items, and correlations between random slopes and intercepts by both participants and items.

All fixed effects were assigned a normal prior with mean 0 and standard deviation of 1.5; intercepts were assigned a normal prior with mean 0 and a standard deviation of 10. These are fairly uninformative priors, as they do not place strong constraints on the model's predictions, and incorporate very little knowledge about what makes a plausible distribution of the dependent measures. We used LKJ(2) as the prior for our correlation matrix. For each model, we ran four Monte Carlo Markov chains in parallel, with 4,000 samples each. The first 2,000 samples of each chain were always discarded as part of warm-up. For all the models reported below, the R-hat statistic was at 1.0, and no divergences were observed.

Our original planned analyses for the Bound and Coreferential model included all trials. However, we observed a very high rate of 'neither' responses in the Pronoun condition. The high rate of 'neither' responses for the Pronoun condition led to significantly lower probabilities of both Coreferential and Bound responses in the original planned models. This makes the effect uninformative about the preferred interpretation of the Pronoun condition, *when that prompt was interpreted as an acceptable continuation of the discourse*. Since one of our predictions critically concerns how the Pronoun is interpreted, we excluded all trials with a 'neither' response in the Bound and Coreferential Model analyses. Thus, the Bound and Coreferential Models reported below quantify the probability of accepting a Bound or Coreferential interpretation of the prompt, conditional on having accepted the prompt as a licit continuation of the discourse. The original models are not reported here, but may be consulted in the associated OSF (see https://osf.io/z6jfh/?view_only=1768a983e5ec4965b403875336f33ee0).

3.6 Results

3.6.1 Analysis of the Form and Neither controls

The descriptive results of our form control conditions are in Figure 10, and the neither control conditions in Figure 11

Responses to the 'neither' controls were overwhelmingly 'neither' responses, ranging from 82% to 89% across conditions. This confirms that participants were generally sensitive to the discourse structure of the text exchanges, and expected that the character mentioned in the critical test sentence would also be mentioned in the missing text message.

Responses to the 'form' controls confirm this conclusion. When the critical test sentence contained a reflexive, or a proper name, participants overwhelmingly chose an inferred context where the prompt subject was mentioned (91% of trials in the *Reflexive* condition, 93% of trials in the *Proper Name* condition). However, when the critical test sentence contained a pronoun, they overwhelmingly chose inferred contexts where the prompt subject was not mentioned (77%). This demonstrates strong sensitivity to the content of the critical test sentence, since as expected, comprehenders overwhelmingly

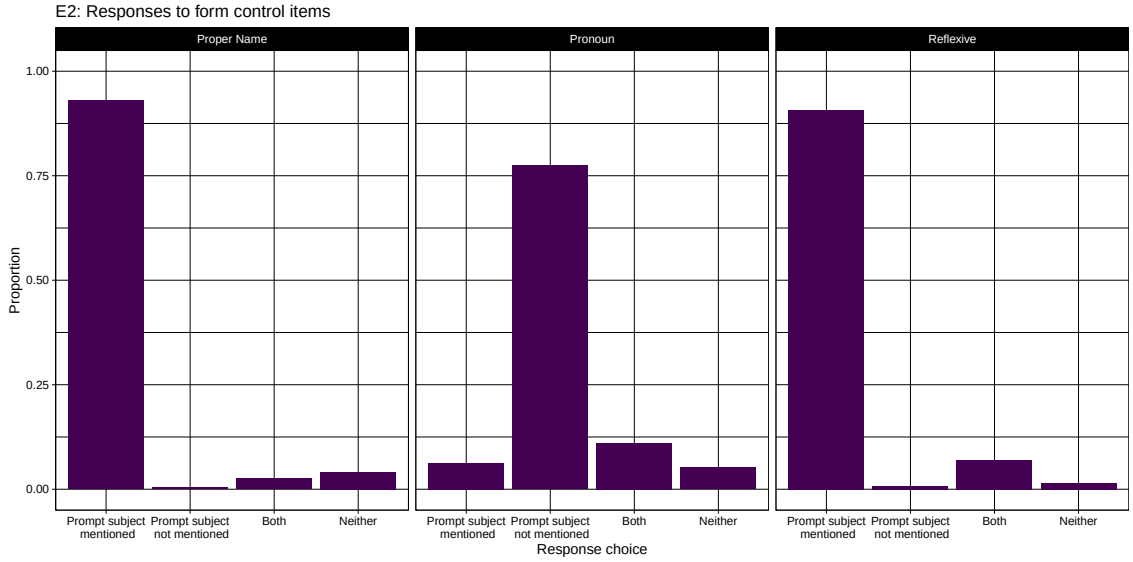

Figure 10: Descriptive summary of response choice behavior for 'Form' control conditions in Experiment 2. Each panel shows the proportion of response choices by each prompt type in the experiment (Proper Name, Pronoun, and Reflexve).

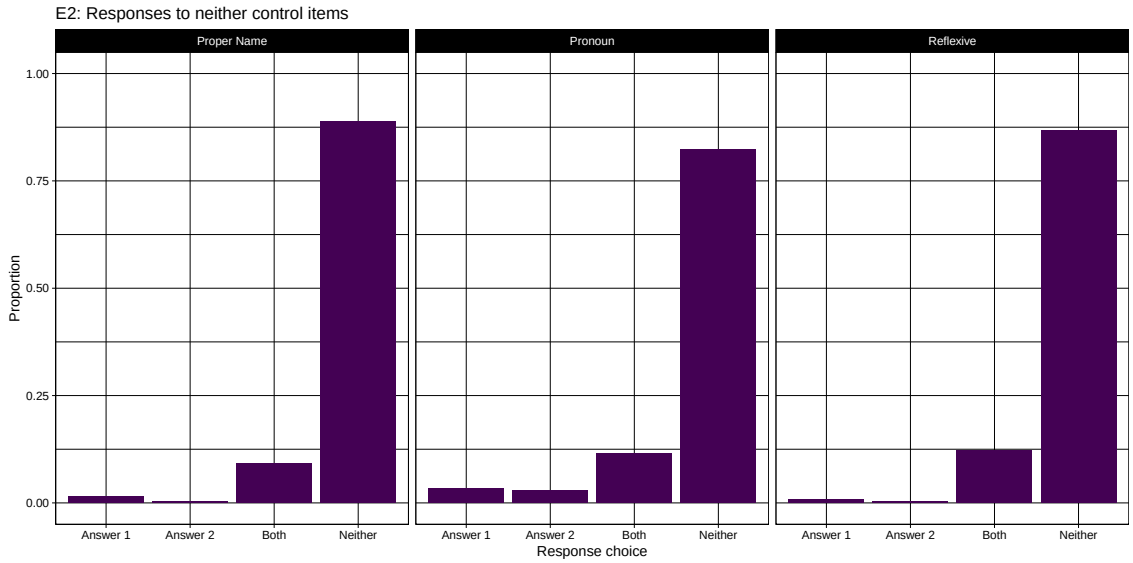

Figure 11: Descriptive summary of response choice behavior for 'Neither' control conditions in Experiment 2. Each panel shows the proportion of response choices by each prompt type in the experiment (Proper Name, Pronoun, and Reflexve).

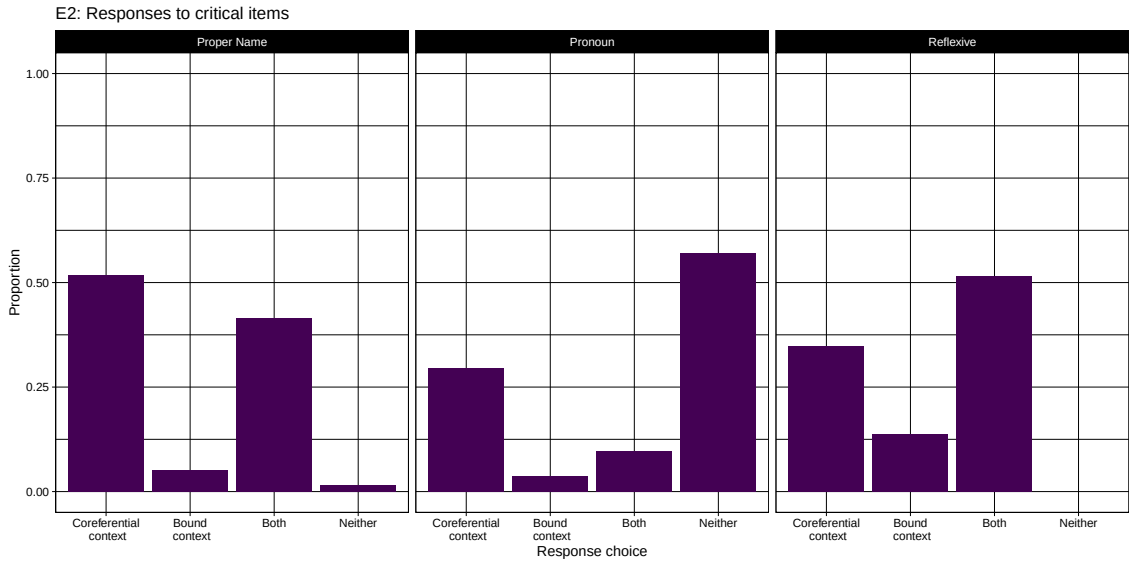

Figure 12: Descriptive summary of response choice behavior for critical trials in Experiment 2. Each panel shows the proportion of response choices by each prompt type in the experiment (Proper Name, Pronoun, and Reflexve).

chose an inferred context that allowed the pronoun to receive a non-locally covalued interpretation.

Having confirmed that comprehenders displayed the expected sensitivity to the structure of the discourse and test sentence in their responses, we turn to the response behavior in the critical trials.

3.6.2 Analysis of critical trials

The raw distribution of responses across prompt type is presented in Figure 12. The posterior estimates over all model parameters are presented in Table 7.

The Neither Model revealed that both the Reflexive and Proper Name conditions had a significantly lower rate of 'neither' responses than the Pronoun condition. The empirical mean rate of 'neither' responses, along with the estimated marginal mean rate of 'neither' responses for each condition from the fitted Neither model, can be seen in Figure 13. The estimated marginal means from this model revealed an essentially zero rate of 'neither' responses for reflexives (95% Credible Interval: [.000,.011]) and proper names (95% Credible Interval: [.002,.023]). In contrast, for pronouns, the estimated marginal mean rate of 'neither' responses was rather high (95% Credible Interval: [.454,.660]), though this was still less than the 82% rate of 'neither' responses in the Pronoun 'neither' controls.

The Coreferential model revealed that pronouns were interpreted as coreferential at a very high rate. The estimated marginal mean rate of 'coreferential' interpretations for pronouns was near ceiling (95% Credible Interval: [.891,.983]). Though the estimated marginal rate of coreferential interpretations for reflexives was slightly lower on average (95% Credible Interval: [.860,.966]), this did not differ significantly from the rate of coreferential interpretations assigned to pronouns. Proper names were given a coreferential interpretation at a higher rate than pronouns (95% Credible Interval: [0.961,.999]).

The Bound model revealed that pronouns were interpreted as bound at a very low rate (95% Credible Interval: [.083,.342]). Both reflexives and names were assigned bound interpretations at significantly higher rates than pronouns, according to this model. The

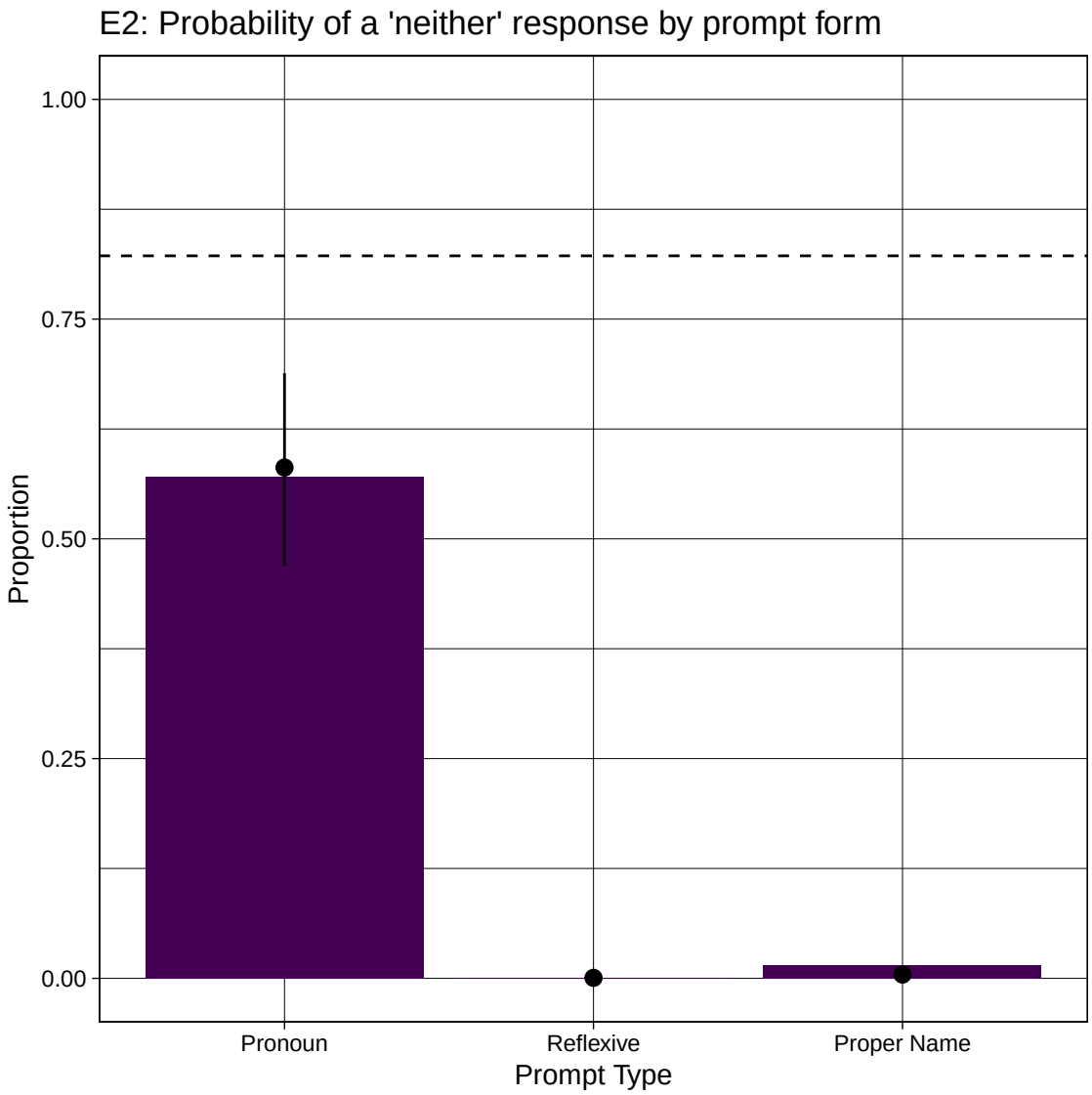

Figure 13: Probability of a ‘Neither’ response across conditions. Bars indicate the mean empirical proportion of Neither responses. Points and error bars indicate the estimated marginal mean and 95% HDI by condition from the fitted Bayesian multilevel logistic regression model. The dashed horizontal line indicates the mean rate of ‘neither’ responses for Pronoun prompts in the Neither Control condition.

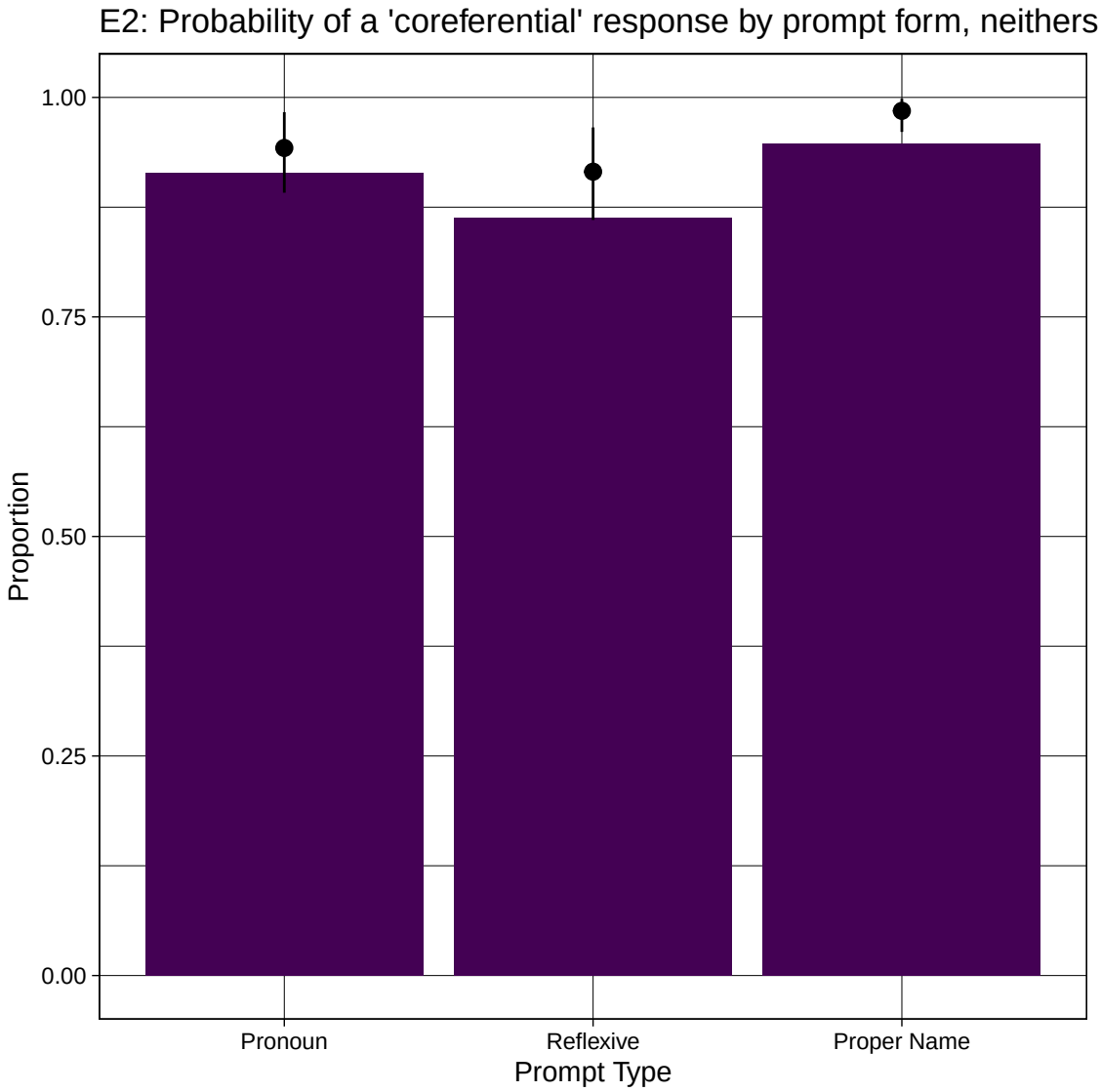

Figure 14: Probability of a Coreferential response across conditions. Bars indicate the mean empirical proportion of Neither responses. Points and error bars indicate the estimated marginal mean and 95% HDI by condition from the fitted Bayesian multilevel logistic regression model.

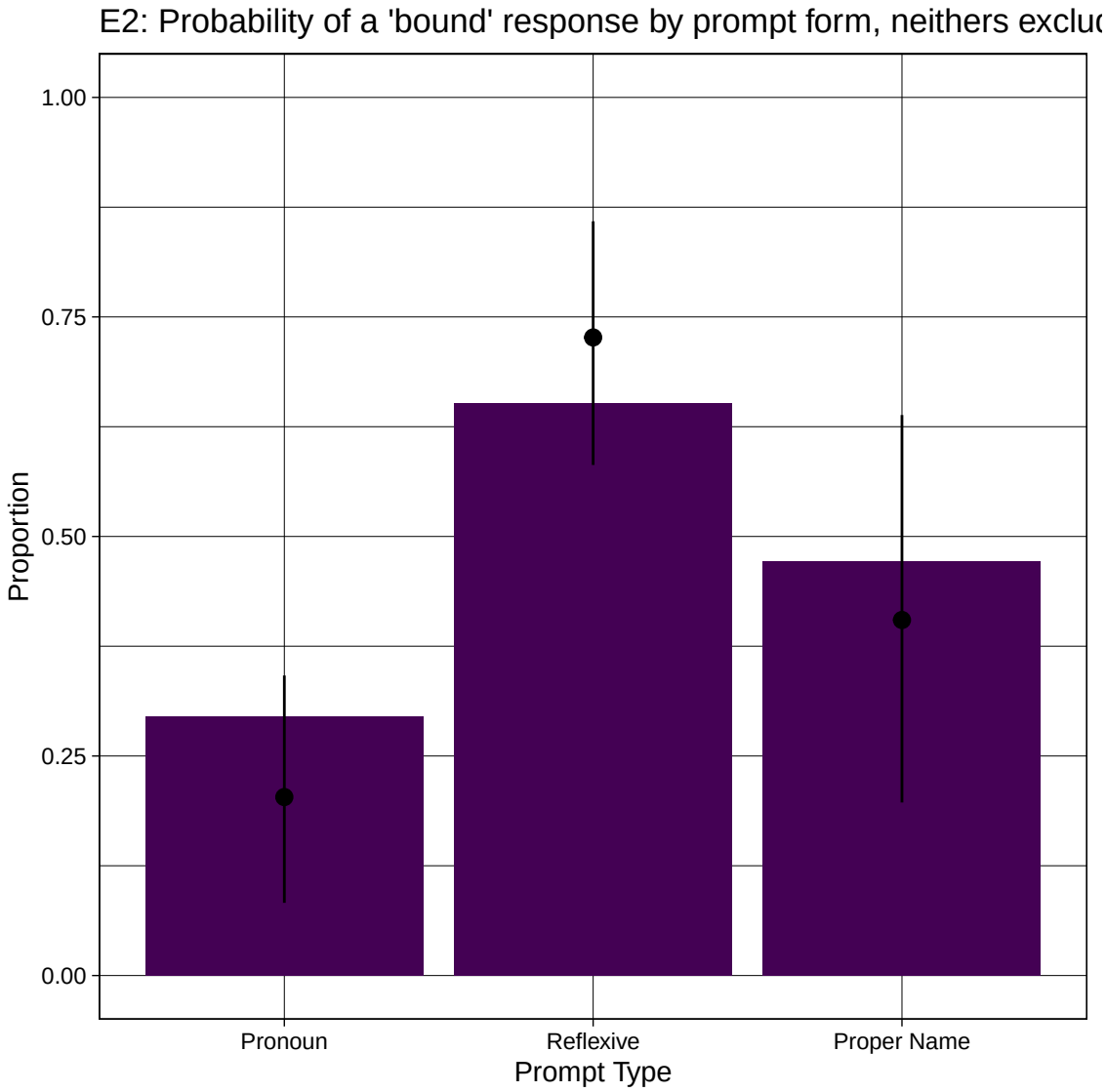

Figure 15: Probability of a Bound response across conditions. Bars indicate the mean empirical proportion of Neither responses. Points and error bars indicate the estimated marginal mean and 95% HDI by condition from the fitted Bayesian multilevel logistic regression model.

Model	Effect	$\hat{\beta}$	95% HPD	
Neither	Intercept (=Pronoun)	.24	[-.19,.66]	Pr(β>0)=.88
	Reflexive vs. Pronoun	-5.80	[-7.33,-4.53]	Pr(β>0)=0
	Proper Name vs. Pronoun	-4.83	[-6.04,-3.83]	Pr(β>0)=0
Coreferential	Intercept (=Pronoun)	2.78	[1.96, 3.85]	Pr(β>0)=1
	Reflexive vs. Pronoun	-0.39	[-1.46, 0.67]	Pr(β>0)=.23
	Proper Name	1.85	[0.21,4.14]	Pr(β>0)=.99
Bound	Intercept (=Pronoun)	-1.55	[-2.53,-0.69]	Pr(β>0)=0
	Reflexive vs. Pronoun	2.62	[1.70,3.65]	Pr(β>0)=1
	Proper Name	1.19	[0.16, 2.22]	Pr(β>0)=.98

Table 7: Posterior estimates for all fixed effects in the Bayesian multilevel logistic regression model of the probability of a response allowing Neither (top), Coreferential (middle) or Bound (bottom) context as a function of the form of the referring expression in the target sentence.

estimated marginal rate of bound interpretations for reflexives was higher than that for names (95% Credible Interval for reflexives: [.581,.859]; 95% Credible Interval for proper names: [.197,.0.638])

3.7 Discussion

In the context selection task, participants largely rejected any context that would allow the pronoun to be locally covalued, reporting that 'neither' context was acceptable between 45 and 66% of the time. This was far higher than the rate at which 'neither' was chosen for either names or reflexives ($<1\%$), but less than the rate at which participants chose 'neither' in the neither control items with a pronoun prompt (82%).

This provides partial confirmation of the prediction of the Semantic Principle B: locally covalued pronouns were generally ruled out no matter the discourse context. However, a straightforward interpretation of this model would lead us to predict that the rate of 'neither' responses for pronoun prompts should be the same as that seen in the neither controls, which we did not see. Instead, we saw a minority of trials in which participants chose another response option, indicating acceptance of a locally covalued interpretation of the pronoun.

On those trials, pronoun prompts were associated with a very high rate of Coreferential responses: between 89 and 98%. However, this rate was not significantly different from the rate at which reflexives were assigned a Coreferential interpretation, and it was lower than the rate of Coreferential interpretations assigned to proper names. Pronouns were also associated with a low, but non-negligible rate of Bound responses, between 8 and 32%, a rate that was lower than both reflexives and proper names.

This latter finding provides partial confirmation of the predictions of a Competition model. Note that this is only partial satisfaction of the Competition View predictions: While pronouns are associated with Coreferential interpretations at a very high rate, we did not see evidence that the rate of coreferential interpretations for pronouns was *higher* than the other response forms. Similarly, while we do find that pronouns have a much lower rate of bound responses than the other forms, as predicted, we do not find that the rate of bound interpretations associated with pronouns is zero, as might have been predicted on a straightforward interpretation of the Competition View.

In brief, we find partial satisfaction of both the Semantic Principle B and Competition views in our context selection task. We turn to a broader interpretation of these observations in light of these different views on local disjoint reference effects in the General Discussion.

4 General discussion

In two experiments, we set out to evaluate to what degree English speakers allow locally covalued pronouns in so-called Evans sentences. These sentences were selected because they have long been thought to distinguish between two broad perspectives on the disjoint reference effect, which we have referred to as the *Semantic Principle B view* and the *Competition view*.

In Experiment 1, we found that producers essentially never produced locally covalued pronouns, even in Coreferential discourse contexts that ruled out bound variable interpretations and reduced the essentially categorical preference for reflexives compared to the Bound Context. In Experiment 2, we found that comprehenders generally rejected any discourse contexts that would force a locally covalued interpretation for a pronoun, including the Coreferential Context where reflexives were seen to be less available in Experiment 1. However, participants in Experiment 2 did not completely reject locally covalued pronouns: on a minority of trials, participants did select a discourse context that permitted a locally covalued pronoun. An analysis of response choice behavior on these trials reveals that comprehenders overwhelmingly preferred a coreferential interpretation of the locally covalued pronoun, rather than the bound variable interpretation, as predicted by Rule I and Grodzinsky & Reinhart (1993).

On balance, these results meet the predictions of both the Semantic Principle B view, and to a lesser extent, the Competition view. Consistent with the Principle B view, there is a strong, contextually-invariant prohibition against locally covalued pronouns in production and in comprehension. However, comprehenders do show the interpretive preferences predicted by the Competition view on the minority of trials where they did permit a locally covalued interpretation of the pronoun.

Overall, this pattern of results suggests that both hypotheses capture complementary aspects of the phenomenon of local disjoint reference. We would like to stress that developing a full formal analysis of this is beyond the scope of this paper, but we think that the present results provide constraints on future analyses that we hope will advance the study of grammatical constraints on reference. We suggest that there are three central theoretical implications of our findings which future analyses might target.

First, *the constraint that rules out local covaluation between its pronoun and antecedent must be semantic (as opposed to syntactic) in nature and apply invariantly across contexts*. In both production and comprehension, locally covalued pronouns are generally neither produced nor considered part of a coherent discourse. This falls out naturally from a semantically non-naive formulation of Principle B (Heim 2007; Jacobson 2007; Schlenker

2005). Such a constraint needs to rule out a syntactic binding mechanism for achieving local covaluation between a pronoun and its local subject, but also needs to rule out *coreference* in this context as well. Importantly, our results show that an unambiguously *coreferential* dependency between a non-reflexive pronoun and a local antecedent is not generally tolerated by English speakers. It is this aspect of the results which lends support to a proposal like Heim/Jacobson/Schlenker's, which directly states the constraint in terms of the syntax-semantics interface. A pragmatic account such as Rule I would struggle to account for the fact that English speakers did not generally tolerate local coreference of pronouns even in discourse contexts where that interpretation differed from a bound variable interpretation.

Second, *comprehenders nonetheless do sometimes interpret locally covalued pronouns in a way that is predicted by a pragmatically-oriented competition theory*. On the minority of trials in Experiment 2 where a locally covalued interpretation was given to the pronoun, bound interpretations of the pronouns were strongly dispreferred. As detailed above, this aligns with the predictions of Rule I. Although Rule I was our point of departure, more generally this observation is compatible with any theory that i) recognizes that reflexive pronouns are the preferred expression of a bound variable dependency with a local subject and ii) proposes that the comprehender can draw inferences about an interpretation based on this fact. We expand on this suggestion below.

Finally, *reflexives can be free variables, in addition to bound.* Subjects in Experiment 1 overwhelmingly chose a reflexive to express coreference. This isn't consistent with the view that reflexives can only be interpreted as bound variables (i.e. (9)). Reflexives, like pronouns, can be free variables. This conclusion has been reached before by observing that in some cases reflexives give rise to strict readings in ellipsis contexts (Hestvik (1992; 1995), McKillen (2016), Bruening (2021)). But this evidence is notoriously difficult to interpret (see Fox (2000) and Fleisher (2023)), and hasn't been decisive. Our findings give independent support for this conjecture.

4.1 Competition in a broader context

If a Semantic Principle B and the type of Competition ascribed to the Competition View are construed as mutually exclusive hypotheses about locally disjoint pronouns, then the first and second theoretical takeaways we describe above create a strong tension. Addressing this tension is a priority for future theoretical work in this area.

One way that this tension might be resolved is by considering competition in a broader context. Rule I is typically considered narrowly as a description of a pragmatic inference with a special application to pronouns in particular grammatical positions. But the role of competition in language interpretation—reasoning over alternatives to an interpretation—has come to be dominated by probabilistic Neo-Gricean perspectives on this process, independently of theories of Principle B (Breheny 2019; Gotzner et al. 2024; Schumacher 2017). For example, the Rational Speech Acts model casts the interpretation process as Bayesian inference driving social reasoning. On this model, a comprehender uses their knowledge of the grammar to simulate how a pragmatically savvy speaker would craft their utterance in context. This, in turn, allows this model to use general reasoning processes to account for a range of ad hoc and scalar implicatures (Degen 2023; Goodman & Frank 2016).

More specifically, the Rational Speech Acts model holds that the in-context interpretation of an utterance could be modeled as the comprehender evaluating the probability of a given interpretation given an utterance. The theory allows for some flexibility in how a

comprehender may evaluate this probability. On this theory, a comprehender may adopt a so-called '0th order' interpretation procedure, where this probability is simply determined by the denotation of a linguistic expression: a world (interpretation) is deemed compatible with an expression iff it is in the set of worlds picked out by that expression's denotation (a *literal listener)*. However, the Rational Speech Acts model posits that listeners may engage in a sort of recursive reasoning, for example, by simulating the outputs of a speaker who is choosing their productions by simulating how a listener would interpret them. Such a *pragmatic listener* will tend to interpret linguistic expressions in a way that respects their *utility* to a cooperative speaker, e.g., the degree to which an expression would unambiguously signal the target meaning to a listener.

By admitting different modes of pragmatic reasoning in this way, a suitably elaborated Rational Speech Acts model may be able to capture our results. If we suppose, following a Semantic Principle B, that the denotation of an expression with a pronoun in object position grammatically excludes any interpretation whereby the pronoun and the subject are locally covalued, then nothing more needs to be said about the behavior of a literal listener in this view: all interpretations incompatible with this denotation will be excluded. But a listener engaging in higher order recursive reasoning about potential speakers could potentially leverage the fact that *if* an English speaker intends to express a locally bound meaning, they will almost categorically do so with a reflexive—as we saw in Experiment 1. If a comprehender is in possession of this fact, then they may be able to exclude the bound variable context for a Pronoun prompt in the context selection task simply by simulating that a speaker would most probably have instead used a reflexive in the target sentence if the bound context was allowed. This would in turn allow them to surmise that among the options given, the bound variable response should be discounted, as it is strongly associated with a different form that was not used in the relevant prompt (the reflexive).

This view might also help explain why intuitive reports on the acceptability of local coreference for pronouns in Evans contexts have been so unstable across analysts in the literature. This model posits that language users have several distinct mechanisms that they might draw upon in forming an intuition (see, e.g., Duff et al. (2025) for a discussion of the pressures that might move comprehenders between different interpretive regimes). For example, they may be subjectively evaluating the probability that they would use a pronoun in a context, which might be extremely low given the existence of a constraint like Principle B. Or they may be evaluating the probability that they can infer a preferred context for a pronoun using second-order pragmatic reasoning. We speculate that such differences in the subjective strategy used to evaluate the suitability of local coreference for pronominals could contribute to the persistent lack of agreement about examples like the Evans examples.

Stepping back, if the account we have sketched here is on the right track, we may characterize the competence component of an English speaker as containing i) a Semantic Principle B and ii) the knowledge that locally covalued bound variable interpretations are overwhelmingly expressed using reflexive pronouns in English. These two bits of grammatical knowledge, combined with general reasoning of the sort captured by the Rational Speech Acts model and related proposals, could well suffice to capture the pattern of results we see. We leave it to future work to formalize this theory and rigorously evaluate it against the data.

4.2 An unexpected finding: Avoidance of local coreference

One unexpected but notable finding from Experiment 1 is the observation that producers tend to *avoid* producing locally coreferential expressions, regardless of the form used to deliver that expression. This was reflected in the reduced proportion of On Target responses for locally bound and (especially) locally coreferential continuations, in comparison to the Non-Coreferential control condition. That is, even though our discourse manipulation was successful in leading participants in Experiment 1 to continue the discourse in the expected fashion (e.g. with a verb phrase that was parallel to the antecedent verb phrase in the context), this pressure was notably weakened in the conditions where ensuring this discourse parallelism required the participant to express a meaning where the subject and object were coreferent.

One deflationary possibility for this finding is that participants simply wanted to see the previously mentioned referent mentioned again in the prompt. In both the Coreferential and Bound context conditions, this referent was indeed mentioned in the subject position. If participants did want to see the previously mentioned referent mentioned again, then in these contexts its presence in the prompt would allow participants to provide more wide-ranging, and hence off target, responses. However, this simple possibility would not account for the difference between Bound and Coreferential contexts: Participants *especially* avoided continuing the text exchange in a discourse parallel way in the Coreferential condition. In other words, it appears that there was a distinct pressure against continuing the discourse in a discourse-parallel way when doing so would require them to express a local coreference relation.

This observation could suggest that there is some degree of ineffability associated with local coreference. This might arise if, for example, all possible grammatical ways of expressing local coreference were to some degree suboptiomal or dispreferred. There's good reason to suspect that this might be the case. As we have argued here, non-reflexive pronouns are subject to a semantic constraint that rules out locally coreferential interpretations. But the alternative formulations may be dispreferred as well, for other reasons.

Consider the three primary alternative formulations we saw in Experiment 1 in (20). By hypothesis, continuing a Coreferential Context discourse with a pronoun leads to a violation of a Semantic Principle B. Another possibility is using a proper name to encode local coreference, but this creates a situation where *Dr. Grant* is repeated twice. This plausibly incurs a repeated name penalty, which has been shown to cause difficulty in processing and leads to lowered judgments of acceptability (Almor & Nair 2007; Gordon et al. 1993; Gordon & Chan 1995; Gordon & Hendrick 1997; 1998; Ledoux et al. 2007); the natural, most economic move would be to refer back with a minimal form, in line with Accessibility Theory (Ariel 1990), according to which highly accessible antecedents are referred to via shorter/minimal forms, while less accessible antecedents are referred to via fuller forms. In addition, this configuration is predicted to be unacceptable given Principle C (Chomsky 1981), which prohibits names from being coreferent with a c-commanding nominal. A final option is to continue this context with a reflexive. While we have shown that reflexives *can* be interpreted coreferentially, our own intuition suggests that this is not the preferred interpretation of a reflexive pronoun. If this intuition is on the right track, then the reflexive continuation of the Coreferential Context may be perceived as less acceptable or natural insofar as it suggests an alternative, locally bound meaning of the target sentence, which is not coherent with the preceding Coreferential Context.

(20) **Context:** Dr. Ellie Sattler said that everyone voted for Dr. Alan Grant, but she lied.

a. Only Dr. Grant voted for him. (*Principle B)
b. Only Dr. Grant voted for Dr. Grant. (*Repeated Name/*Principle C)
c. Only Dr. Grant voted for himself. (*Discourse coherence)

If our speculative interpretation of this effect is on the right track, then this provides an additional, albeit unexpected, measure of support for a Semantic Principle B; even when the discourse strongly pushes them to do so, producers appear to choose to continue the discourse in a non-parallel fashion rather than linguistically encode local coreference, be it with a pronoun or otherwise. Given this, this unexpected finding seems to suggest that producers even rank violating expected discourse parallelism/coherence above using a pronoun to express local coreference, an unexpected but perhaps strong point in favor of a Semantic Principle B.

5 Conclusion

In two experiments, we tested the degree to which English producers and comprehenders would allow a non-reflexive pronoun to express local coreference. We found that English speakers behave as if there is a general, discourse invariant prohibition against a non-reflexive pronoun expressing local coreference. This finding suggests that the correct formulation of Principle B should rule out both locally bound and coreferential interpretations of pronouns, as in semantically explicit formulations of Principle B Heim (2007); Jacobson (2007); Schlenker (2005). However, we did find that comprehenders nonetheless do sometimes exhibit a preference for coreferential over bound variable interpretations of locally covalued pronouns in a context selection task. This suggests that competitive interpretation mechanisms can nonetheless guide the interpretation of locally covalued pronouns in some contexts. We suggested that these apparently conflicting observations can be reconciled by augmenting a speaker guided by an invariant Principle B with a Gricean pragmatic reasoning mechanism that can be flexibly deployed in different task contexts.

Data availability

The experimental items, deidentified data, the visualization and analysis scripts, and other supplementary files associated with this article are openly available in Open Science Framework: https://osf.io/z6jfh/?view_only=1768a983e5ec4965b403875336f33ee0.

Ethics and consent

These experiments were run under IRB Protocol #2929 ('Internet-based psycholinguistic questionnaires') approved by the UMass Institutional Review Board.

Funding information

This work was supported by NSF BCS #1941485 to Brian Dillon. During the revision of the current study, A.C. Bleotu was partly supported by the Austrian Science Fund (FWF) 10.55776/F1003 within the SFB project *Language between Redundancy and Deficiency*.

Acknowledgements

We are grateful to feedback and support from the members of the Binding Theory at UMass Reading Group over the course of this project. We would also like to acknowledge helpful discussion and feedback from Dorothy Ahn, Shannon Bryant, Ivy Sichel, Ken Safir, and Kristen Syrett, as well as from audience members at Rutgers University Department of Linguistics, and the University of Wisconsin-Madison's Department of Language Sciences. All errors remain our own.

Competing interests

The authors have no competing interests to declare.

CRediT authorship contribution statement

Breanna Pratley: Conceptualization, Data curation, Investigation, Methodology, Writing (original draft), Writing (review & editing), **Jed Sam Pizarro-Guevara:** Conceptualization, Investigation, Methodology, Data curation, Formal analyses, Visualization, Writing (review & editing),**Adina Camelia Bleotu:** Conceptualization, Investigation, Methodology, Writing (review & editing), **Kyle Johnson:** Conceptualization, Investigation, Methodology, Writing (review & editing), **Brian Dillon:** Conceptualization, Data curation, Formal analyses, Funding Acquisition, Project administration, Visualization, Writing (review & editing).

References

Almor, Amit & Nair, Veena A. 2007. The form of referential expressions in discourse. *Language and Linguistics Compass* 1(1-2). 84–99.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Avrutin, S. & Wexler, Kenneth. 1992. Development of principle b in russian: Coindexation at lf and coreference. *Language Acquisition* 2. 259–306.

Baauw, Sergio M. & Escobar, Linda & Philip, William. 1997. A delay of principle b effect in spanish speaking children: The role of lexical feature acquisition. In Sorace, Antonella & Heycock, Caroline & Shillcock, Robin (eds.), *Proceedings of the gala 97 conference on language acquisition*. 16–21. Edinburgh: University of Edinburgh.

Baauw, Susan. 2013. *Grammatical features and the acquisition of reference: A comparative study of dutch and spanish*. New York: Routledge.

Barker, Chris. 2012. Quantificational binding does not require c-command. *Linguistic Inquiry* 43(4). 614–633.

Bassel, Noa. 2024. *Complex anaphors*: Hebrew University of Jerusalem dissertation.

Breheny, Richard. 2019. 39scalar implicatures. In *The oxford handbook of experimental semantics and pragmatics*, Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198791768.013.4. https://doi.org/10.1093/oxfordhb/9780198791768.013.4

Bruening, Benjamin. 2021. Generalizing the presuppositional approach to the binding conditions. *Syntax* 24(4). 417–461. https://doi.org/10.1111/synt.12221. http://doi.org/10.1111/synt.12221

Bürkner, Paul-Christian. 2017. brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–80. 10.18637/jss.v080.i01.

Charnavel, Isabelle. 2019. *Locality and logophoricity: A theory of exempt anaphora*. Oxford University Press.

Chien, Yu-Chin & Wexler, Ken. 1987. Children's acquisition of reflexives and pronouns. *Papers and Reports on Child Language Development* 26. 30–39.

Chien, Yu-chin & Wexler, Kenneth. 1990. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition: A Journal of Developmental Linguistics* 1(3). 225–295. https://doi.org/10.1207/s15327817la0103_2. https://doi.org/10.1207/s15327817la0103_2

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Conroy, Anastasia & Takahashi, Eri & Lidz, Jeffrey & Phillips, Colin. 2009. Equal treatment for all antecedents: How children succeed with principle b. *Linguistic Inquiry* 40(3). 446–486. https://doi.org/10.1162/ling.2009.40.3.446. https://doi.org/10.1162/ling.2009.40.3.446

Degen, Judith. 2023. The rational speech act framework. *Annual Review of Linguistics* 9(1). 519–540.

Dowty, David R. 1980. Comments on the paper by Bach and Partee. In Kreiman, Jody & Ojeda, Almerindo K. (eds.), *Papers from the parasesion on pronouns and anaphora*. 29–40. Chicago Linguistics Society.

Duff, John & Mayn, Alexandra & Demberg, Vera. 2025. An act-r model of resource-rational performance in a pragmatic reference game. In *Proceedings of the annual meeting of the cognitive science society*, vol. 47.

Elbourne, Paul. 2005. On the acquisition of principle b. *Linguistic Inquiry* 36(3). 333–365. http://www.jstor.org/stable/4179328.

Evans, Gareth. 1980. Pronouns. *Linguistic Inquiry* 11. 337–362.

Fleisher, Nicholas. 2023. On referential parallelism and compulsory binding. *Linguistic Inquiry* 54(4). 841–860.

Fox, Danny. 2000. *Economy and semantic interpretation*. Cambridge, Massachusetts: MIT Press.

Gamut, L. T. F. 1990. *Logic, language and meaning vol. 1*. University of Chicago Press.

Gelman, Andrew & Lee, Daniel & Guo, Jiqiang. 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* 40(5). 530–43.

Goodman, Noah D & Frank, Michael C. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* 20(11). 818–829.

Gordon, Peter C & Chan, Davina. 1995. Pronouns, passives, and discourse coherence. *Journal of Memory and Language* 34(2). 216–231.

Gordon, Peter C & Grosz, Barbara J & Gilliom, Laura A. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive science* 17(3). 311–347.

Gordon, Peter C & Hendrick, Randall. 1997. Intuitive knowledge of linguistic coreference. *Cognition* 62(3). 325–370.

Gordon, Peter C & Hendrick, Randall. 1998. The representation and processing of coreference in discourse. *Cognitive science* 22(4). 389–424.

Gotzner, Nicole & Harris, Jesse A & Breheny, Richard & Sharvit, Yael. 2024. *Alternatives in grammar and cognition*. Springer.

Grodzinsky, Yosef & Reinhart, Tanya. 1993. The innateness of binding and coreference. *Linguistic Inquiry* 24. 69–102. https://api.semanticscholar.org/CorpusID:17347569.

Hamann, Cornelia & Kowalski, Odette & Philip, William. 1997. The french "delay of principle b" effect. In Hughes, Elizabeth & Hughes, Mary & Greenhill, Annabel (eds.), *Bucld 21 proceedings*. 205–219. Somerville, MA: Cascadilla Press.

Heim, Irene. 2007. Forks in the road to rule i. In Abdurrahman, M. & Schardl, A. & Walkow, M. (eds.), *Nels 38: Proceedings of the thirty-eighth annual meeting of the north east linguistic society*.

Heim, Irene & Kratzer, Angelika. 1998. *Semantics in generative grammar*. Malden, MA: Blackwell.

Hendriks, Petra & Spenader, Jennifer. 2006. When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language acquisition* 13(4). 319–348.

Hestvik, Arild. 1992. Subordination and strict identity of interpretation of reflexives. In Berman, Steve & Hestvik, Arild (eds.), *Proceedings of the stuttgart ellipsis workshop*. Stuttgart.

Hestvik, Arild. 1995. Reflexives and ellipsis. *Natural Language Semantics* 3(2). 211–237.

Jacobson, Pauline. 2007. Direct compositionality and variable-free semantics: The case of "principle b" effects. In *Direct compositionality*, Oxford University Press. https://doi.org/10.1093/oso/9780199204373.003.0006. https://doi.org/10.1093/oso/9780199204373.003.0006

Keenan, Edward L. 1971. Quantifier structures in english. *Foundations of language* 255–284.

Kiparsky, Paul. 2002. Disjoint reference and the typology of. *More than words: A festschrift for Dieter Wunderlich* 53. 179.

Klobučar, Nina & Folli, Raffaella & Sevdali, Christina & Gerard, John. 2025. (all) pronouns are difficult, but not delayed. In *Proceedings of the 49th annual boston university conference on language development*. 370–381. Somerville, MA: Cascadilla Press.

Kroll, Margaret. 2020. *Comprehending ellipsis*: University of California, Santa Cruz dissertation.

Kruschke, John. 2014. *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Kuijper, Sanne JM & Hartman, Catharina A & Hendriks, Petra. 2021. Children's pronoun interpretation problems are related to theory of mind and inhibition, but not working memory. *Frontiers in psychology* 12. 610401.

Ledoux, Kerry & Gordon, Peter C & Camblin, C Christine & Swaab, Tamara Y. 2007. Coreference and lexical repetition: Mechanisms of discourse integration. *Memory & cognition* 35(4). 801–815.

Lenth, Russell V. 2025. *emmeans: Estimated marginal means, aka least-squares means*. https://rvlenth.github.io/emmeans/. R package version 1.11.2-8.

Levinson, Stephen C. 1987. Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena1. *Journal of linguistics* 23(2). 379–434.

Levinson, Stephen C. 1991. Pragmatic reduction of the binding conditions revisited. *Journal of linguistics* 27(1). 107–161.

May, Robert. 1977. *The grammar of quantification*: Massachusetts Institute of Technology dissertation.

May, Robert. 1985. *Logical form: Its structure and derivation*. Cambridge, Massachusetts: MIT Press.

McKee, Simon. 1992. A comparison of pronouns and anaphors in italian and english acquisition. *Language Acquisition* 2. 21–54.

McKillen, Alanah. 2016. *On the interpretation of reflexive pronouns*: McGill University dissertation.

Philip, William & Coopmans, Peter. 1996. The double dutch delay of principle b effect. In Stringfellow, Amanda & Cahana-Amitay, Dalia & Hughes, Elaine & Zukowski, Andrea

(eds.), *Proceedings of the 20th boston university conference on language development*. 576–587. Somerville: Cascadilla Press.

Pinto, Manuela & Zuckerman, Shalom. 2018. Coloring book: A new method for testing language comprehension. *Behavior Research Methods* 1–20. https://doi.org/10.3758/s13428-018-1114-8. https://doi.org/10.3758/s13428-018-1114-8

Pollard, Carl & Sag, Ivan A. 1992. Anaphors in english and the scope of binding theory. *Linguistic inquiry* 23(2). 261–303.

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Reinhart, Tanya. 1983. Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy* 6(1). 47–88. https://doi.org/10.1007/bf00868090

Reinhart, Tanya. 2006. *Interface strategies: Optimal and costly computations* (Linguistic Inquiry Monograph 45). Cambridge, MA: MIT Press.

Roelofsen, Floris. 2010. Condition b effects in two simple steps. *Natural Language Semantics* 18. 115–140. https://doi.org/10.1007/s11050-009-9049-3. https://doi.org/10.1007/s11050-009-9049-3

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 117–121.

Safir, Ken. 2014. One true anaphor. *Linguistic Inquiry* 45(1). 91–124.

Safir, Kenneth J. 2004. *The syntax of anaphora*. Oxford University Press.

Sag, Ivan A. 1976. *Deletion and logical form.*: Massachusetts Institute of Technology dissertation.

Schad, Daniel J & Betancourt, Michael & Vasishth, Shravan. 2021. Toward a principled bayesian workflow in cognitive science. *Psychological methods* 26(1). 103.

Schlenker, Philippe. 2005. Non-redundancy: Towards a semantic reinterpretation of binding theory. *Natural Language Semantics* 13(1). 1–92.

Schumacher, Petra B. 2017. Semantic-pragmatic processing. *The handbook of psycholinguistics* 392–410.

Schwarzschild, Roger. 1999. Givenness, avoid F and other constraints on the placement of accent. *Natural Language Semantics* 7(2). 141–177.

Sciullo, Anna Maria Di & Agüero-Bautista, Calixto. 2008. The delay of principle b effect (dpbe) and its absence in some languages. *Language and Speech* 51(1 & 2). 77–100. https://doi.org/10.1177/00238309080510010601

Spenader, Jennifer & Smits, Erik-Jan & Hendriks, Petra. 2009. Coherent discourse solves the pronoun interpretation problem. *Journal of child language* 36(1). 23–52.

Thornton, Rosalind & Wexler, Kenneth. 1999. *Principle b, vp ellipsis, and interpretation in child grammar*. MIT press.

Van Rij, Jacolien & Van Rijn, Hedderik & Hendriks, Petra. 2010. Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language* 37(3). 731–766.

Verbuk, Anna & Roeper, Thomas. 2010. How pragmatics and syntax make principle b acquirable. *Language Acquisition* 17(1-2). 51–65.

Williams, Edwin S. 1977. Discourse and logical form. *Linguistic inquiry* 101–139.

Wyngaerd, Guido Vanden & Rooryck, Johan. 2011. *Dissolving binding theory*. Oxford University: Oxford University Press.