Cognition xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Shared syntax between comprehension and production: Multi-paradigm evidence that resumptive pronouns hinder comprehension☆

Adam M. Morgan[a,*], Titus von der Malsburg[b,c], Victor S. Ferreira[d], Eva Wittenberg[e]

[a] NYU School of Medicine, Department of Neurology, 227 E 30th St, New York, NY 10016, USA
[b] University of Potsdam, Department of Linguistics, Haus 14, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany
[c] MIT, Brain and Cognitive Sciences, 77 Massachusetts Ave, Room 46-2005, Cambridge, MA 02139-4307, USA
[d] UC San Diego, Department of Psychology, 9500 Gilman Dr., La Jolla, CA 92093, USA
[e] UC San Diego, Department of Linguistics, 9500 Gilman Dr., La Jolla, CA 92093, USA

ARTICLE INFO

Keywords:
Syntax
Resumptive pronouns
Language comprehension
Language production
Eyetracking
Multi-paradigm self-replication

ABSTRACT

Language comprehension and production are generally assumed to use the same representations, but *resumption* poses a problem for this view: This structure is regularly produced, but judged highly unacceptable. Production-based solutions to this paradox explain resumption in terms of processing pressures, whereas the *Facilitation Hypothesis* suggests resumption is produced to help listeners comprehend. Previous research purported to support the Facilitation Hypothesis did not test its keystone prediction: that resumption improves *accuracy* of interpretation. Here, we test this prediction directly, controlling for factors that previous work did not. Results show that resumption in fact *hinders* comprehension in the same sentences that native speakers produced, a finding which replicated across four high-powered experiments with varying paradigms: sentence-picture matching ($N = 300$), self-paced reading ($N = 96$), visual world eye-tracking ($N = 96$), and multiple-choice comprehension question ($N = 150$). These findings are consistent with production-based accounts, indicating that comprehension and production may indeed share representations, although our findings point toward a limit on the degree of overlap. Methodologically speaking, the findings highlight the importance of measuring interpretation when studying comprehension.

1. Introduction

Early on in the neuroscience of language comprehension, researchers stumbled upon an intriguing discovery. Humans tend to make silent articulatory movements as if they were producing sentences while comprehending them. While "subvocalization" was initially regarded as a mild nuisance because it created artifacts in fMRI data, it wound up being a clue to something deeper about the architecture of language. More recent research has shown that simply listening to speech evokes activity in motor cortex analogous to the activity observed during speech production (Watkins, Strafella, & Paus, 2003; Wilson, Saygin, Sereno, & Iacoboni, 2004), and interrupting activity in these areas (e.g., with transcranial magnetic stimulation, D'Ausilio et al., 2009) impairs speech perception. Language production seems to be intricately involved in language comprehension.

Exactly how much is shared between the two modalities is a matter of debate. More extreme accounts contend that comprehension relies entirely on production. For instance, *Analysis by Synthesis* approaches hold that in order to comprehend a sentence, a comprehender synthesizes a string to match the input using the production system (see Bever & Poeppel, 2010 for a review). According to this view, the comprehension system at least in part *is* the production system.

The opposite view – that comprehension and production are entirely separable – is generally discounted a priori. It would indeed be quite impractical for the two systems to not share any resources. For instance, if lexical representations were not shared, then there would need to be two separate, redundant lexicons: one for production and one for comprehension. If this were the underlying architecture, how could the system guarantee that these two lexicons are the same – that is, that individuals speak the same language they comprehend?

☆ This research was supported in part by the National Institutes of Health under grant R01-HD051030 to Victor S. Ferreira and by the National Science Foundation under grant DGE-1144086 to Adam M. Morgan. Titus von der Malsburg was supported by a Feodor-Lynen research fellowship awarded by the Alexander von Humboldt Foundation.

* Corresponding author.
E-mail addresses: adam.morgan@nyulangone.org (A.M. Morgan), malsburg@uni-potsdam.de (T. von der Malsburg), vferreira@ucsd.edu (V.S. Ferreira), ewittenberg@ucsd.edu (E. Wittenberg).

https://doi.org/10.1016/j.cognition.2020.104417
Received 4 July 2019; Received in revised form 20 May 2020; Accepted 24 July 2020

There is, however, a particular syntactic structure with properties so puzzling that some researchers have proposed exactly such a system. Imagine that you read a word, perhaps *resumption*, and you don't know what it means. Consider how you might complete the following sentence: "I just read a word that I don't know what...." More often than not, native English speakers would say something like "…it means," rendering the sentence in (1):

(1) I just read a word that I don't know what **it** means.

Resumption, or the use of *resumptive pronouns* like the "it" in (1), poses a problem for standard views about production, comprehension, and grammar. Specifically, acceptability judgments and production, both common metrics for grammaticality, nearly always align: Speakers produce the same kinds of sentences that they find acceptable. However, English resumption does not fit this pattern.

Resumptive pronouns are commonly produced in English, suggesting they are grammatical (Cann, Kaplan, & Kempson, 2005). Examples abound in corpora (Bennett, 2009) and in natural speech:

(2) "We have these things called aircraft carriers, where planes land on **them**."
-Barack Obama (Davidson Sorkin, 2012)

(3) "...the sale of the uranium that nobody knows what **it** means."
-Donald Trump (Noble, 2017)

In experimental settings, resumptive pronouns can be reliably elicited, both in speech (F. Ferreira & Swets, 2005) and writing (Morgan & Wagers, 2018), and both when speakers are under time pressure to respond and when they are not (F. Ferreira & Swets, 2005).

But in comprehension, resumptive pronouns are highly unacceptable, suggesting they are *not* grammatical (Alexopoulou & Keller, 2007; Han et al., 2012; Heestand, Xiang, & Polinsky, 2011; Keffala & Goodall, 2011; Polinsky, Clemens, Morgan, Xiang, & Heestand, 2013). This is true across a wide variety of sentence types, both with written and auditory stimulus presentation (Clemens, Morgan, Polinsky, & Xiang, 2012; Heestand, Xiang, & Polinsky, 2011). It is even true when the comprehender is the same person who just produced the resumptive pronoun (Zukowski & Larsen, 2004).

Resumptive pronouns, then, present a case where the two most common metrics for grammaticality dissociate. This paradox is one of the most intriguing puzzles in the field because it seems to indicate that the standard assumption that comprehension and production share representations is in fact flawed. Indeed, in response to this paradox, F. Ferreira & Swets, 2005 advocate exactly the opposite: "The two systems [production and comprehension] do not consult the exact same database of grammatical rules, as indicated by the finding that the production system allows [resumptive pronouns], but the comprehension system tends to reject them."

But rather than rejecting the notion of shared representations, the paradox may alternatively be resolved if resumptive pronouns are shown to somehow be an exception to the rule. One prominent account along these lines, the *Facilitation Hypothesis*, argues that resumptive pronouns are indeed ungrammatical, but speakers nonetheless produce them because they help listeners keep track of reference. However, in four experiments, we demonstrate that when this hypothesis is tested directly, the data instead indicate that resumptive pronouns *hinder* comprehension. This suggests that the Facilitation Hypothesis should be rejected in favor of production-based models of resumption, which explain resumption as the result of production processes gone awry.

In the following section, we give a brief theoretical overview of resumptive pronouns and discuss prominent approaches to resolving the comprehension-production paradox, with particular focus on the Facilitation Hypothesis. We then argue that to adequately test the Facilitation Hypothesis one must investigate how resumptive pronouns are interpreted, which has yet to be done. We do this in four experiments, which consistently provide evidence against the Facilitation Hypothesis, but consistent with production accounts of resumption. We conclude by discussing the implications for the relationship between comprehension and production.

1.1. Resumptive pronouns in English

Resumptive pronouns do not exist in isolation, but are parts of larger structures, such as *relative clauses*. In (4), for example, "that the fairies kidnapped in the night" is a relative clause that modifies "girl":

(4) the girl [that the fairies kidnapped _ in the night].

The modified noun *'girl,'* or the *head noun*, is not repeated inside the relative clause. The resulting empty position is referred to as a *gap* (indicated with underscores throughout). Structures like relative clauses, where the meaning of a gap corresponds to that of a faraway head noun, are known as *wh-dependencies*. Throughout this paper, we will use relative clauses (more specifically, clefts) to create wh-dependencies.

In English, leaving a gap is the only grammatical way to form wh-dependencies. In contrast, other languages employ resumption. Irish (McCloskey, 2002), Hebrew (Shlonsky, 1992), Gbadi (Koopman, 1983), and Cantonese (Lau, 2016), for instance, allow speakers the option of inserting a resumptive pronoun, as in (1), (2), (3) and (5):

(5) the girl [that the fairies kidnapped her in the night].

Unlike ordinary pronouns, which may refer to any number of potential referents in any language, resumptive pronouns must always refer to the head noun (e.g., Zaenen, Engdahl, & Maling, 1981). So, in (4) and (5), the object of "kidnap" must be "the girl."

Resumptive pronouns are produced more often in *islands*, a class of structures which are unacceptable when a gap appears inside them. For instance, (4) is a *non-island*, meaning that it is acceptable with a gap in it. English speakers produce fewer than 5% resumptive pronouns in non-islands (Morgan & Wagers, 2018). However, in islands like (6) and (7), resumptive pronouns are much more common.

(6) the fairies [that I wonder why _ kidnapped the girl].

(7) the fairies [that I was scared because _ kidnapped the girl].

Island violations like these vary in their degree of unacceptability (Ross, 1967). Example (6) is a *weak island*, and is only moderately unacceptable with a gap, while (7) is a *strong island* and is highly unacceptable with a gap.[1] In a paired comprehension-production study, Morgan & Wagers (2018) showed that English speakers produce resumptive pronouns in a given structure at a rate that correlates strongly with that structure's degree of unacceptability. Thus, in their study, English speakers produced close to 50% resumptive pronouns in weak islands, and over 90% in strong islands.

1.2. Accounting for the comprehension-production paradox

Here, we focus on two of the dominant approaches to the comprehension-production paradox: the Facilitation Hypothesis and production-based accounts. The Facilitation Hypothesis posits that resumptive pronouns are easier to comprehend than gaps (e.g., Beltrama & Xiang,

[1] *Weak* and *strong* islands are actually classes of island structures, each composed of many different structures. In this study, we will use structures called *wh-islands* (as in 6) to operationalize weak islands, and *adjunct islands* (as in 7) as strong islands. For the sake of readability, we use the more intuitive labels *weak island* and *strong island* throughout.

2016; Dickey, 1996; Hofmeister & Norcliffe, 2013; Polinsky, Clemens, Morgan, Xiang, & Heestand, 2013; Prince, 1990). It is easy to imagine why this might be the case. Resumptive pronouns provide an overt cue for the end of a wh-dependency, along with number, gender, and animacy information which might be helpful for retrieving the correct antecedent (F. Ferreira & Swets, 2005). Gaps, by comparison, do none of this. If the Facilitation Hypothesis is true, it is of course puzzling why resumptive pronouns would be ungrammatical in the first place when they are helpful.

To test the Facilitation Hypothesis, Hofmeister and Norcliffe, 2013 presented subjects with sentences that had gaps or resumptive pronouns, as in (8a). Participants read the sentences word by word in a moving window self-paced reading paradigm, and then answered a comprehension question (8b).

(8) a. The prison officials had acknowledged that there was a prisoner that the guard helped _/him to make a daring escape.
b. Was a prisoner able to escape?

Hofmeister and Norcliffe's data revealed that the words immediately following resumptive pronouns were read faster than those following gaps in otherwise identical sentences. They interpreted this result as reflecting "more efficient processing," a sign that "the resumptive pronoun facilitates processing compared to a gap" – a surprising finding given that ungrammatical words usually cause significant slowdowns.

Faster reading times, however, do not necessarily imply facilitated comprehension. Indeed, a number of possibilities are compatible with this pattern of data. One is that gap and resumptive pronoun dependencies are equally easy or difficult to process, but because resumptive pronouns take longer to read than gaps, they spread the same amount of information across more words. If, as some evidence suggests, the system aims to process a uniform amount of information per unit time (Jaeger & Levy, 2007), then Hofmeister and Norcliffe (2013) subjects may have sped up after resumptive pronouns because they had to process less information per word, not because the resumptive pronoun made parsing easier. Indeed, can faster reading times on individual words constitute facilitation if it takes longer to read the whole sentence?

Another possibility, which we will return to throughout the paper, is that readers are simply confused by resumptive pronouns. The decrease in reading times may reflect giving up on parsing and clicking through to end the trial. Given that resumptive pronouns are rare in English, and particularly in non-island contexts like those Hofmeister and Norcliffe tested, this seems like a more likely interpretation of their results than facilitation. (See F. Ferreira, Bailey, & Ferraro, 2002; Nicenboim, Logačev, Gattei, & Vasishth, 2016 for related proposals.)

In order to infer that the faster reading times after resumptive pronouns reflect facilitation, one would need to minimally establish that resumptive pronouns are interpreted at least as correctly as gaps. That is, in order to evaluate the usefulness of resumption in comprehension, one must also measure *interpretation*.

Hofmeister and Norcliffe (2013) did not report the interpretation data collected from comprehension questions, but they did remove data from trials that were interpreted incorrectly before performing their analysis on reading times. On the surface, this would seem to ensure that faster reading times were measured only on correctly answered trials. But their stimuli were pragmatically rich, and as such it may have been possible for participants to correctly answer comprehension questions based on lexical content alone. Just reading the words *prisoner, prison guard, help*, and *escape* can conjure up a plausible scenario, without needing process the syntax of the sentence (see Mollica et al., 2018 for neural evidence that adjacent words are processed compositionally even when they do not form grammatical strings).

If this is the case, then it may be even more likely that faster reading times after resumptive pronouns reflect giving up on parsing and not facilitated parsing. Without more information about exactly how subjects parsed – or assigned a grammatical structure to – sentences, Hofmeister and Norcliffe's data cannot definitively answer the question of whether resumptive pronouns facilitate comprehension.

Beltrama and Xiang (2016) also tested the Facilitation Hypothesis by asking subjects to rate sentences for comprehensibility. Their stimuli consisted of a context paragraph (9a) followed by a target sentence, which were manipulated to appear with gaps or resumptive pronouns in non-islands (9b) or islands (9c).

(9) a. Have you heard? Yesterday there were riots in the streets. Some people were wounded. Look here, they're talking about it in the paper.
b. This is the boy that the cop who was leading the operation beat _/him up.
c. This is the boy that the cop who beat _/him up was leading the operation.

They found that in non-islands, gaps were rated as more comprehensible than resumptive pronouns.[2] In islands, on the other hand, resumptive pronouns were rated as more comprehensible than gaps. Beltrama and Xiang took these results to be consistent with a modified version of the Facilitation Hypothesis: that resumptive pronouns facilitate processing, but only in islands. However, like Hofmeister and Norcliffe (2013), Beltrama and Xiang (2016) used stimuli that provided readers with heavy pragmatic cues and did not report how their participants interpreted them.

Like with reading times, comprehensibility ratings alone are not sufficient. It is in principle possible that resumptive pronouns lead comprehenders to *interpret* sentences less correctly, but to nonetheless *feel* that they are interpreting them more correctly. Knowing how participants interpret resumptive pronouns is therefore crucial.

What remains to be tested in this literature is the keystone prediction of the Facilitation Hypothesis: that resumptive pronouns result in more accurate interpretation than gaps. For a system whose goal is communication, the worst possible outcome is incorrect interpretation. One might even consider decreased processing speeds facilitatory if they corresponded to an increase in correct interpretation. But a decrease in interpretation accuracy can never constitute facilitation. If resumptive pronouns make the listener less likely to understand the intended meaning, then it hardly matters whether they do so in less time or with more confidence. Interestingly however, a decrease in interpretation accuracy is exactly the prediction of the second family of explanations of the comprehension-production paradox.

Production-based theories attempt to resolve the paradox by assuming that resumptive pronouns are ungrammatical, straightforwardly accounting for their unacceptability, and explaining their production in terms of difficulties in online production processes – either pressure to produce locally licit structures (Asudeh, 2004), or a breakdown in dependency maintenance (Morgan & Wagers, 2018). If resumptive pronouns are indeed ungrammatical, then by definition it means that comprehenders cannot parse resumptive dependencies. But when there is no grammatical structure, comprehension should be impaired. Thus, both Asudeh (2004, 2011) and Morgan & Wagers (2018) predict that resumptive pronouns should lead to *worse* comprehension than gaps – the opposite prediction of the Facilitation Hypothesis.

1.3. The present study

There is a growing body of data which, on the surface at least, seems

[2] Note that Hofmeister and Norcliffe (2013) stimuli were all non-islands, but they came to the opposite conclusion on the basis of faster reading times. This discrepancy may trace back to a flaw in the common simplifying assumption that faster reading times mean more efficient processing.

to support the Facilitation Hypothesis (Beltrama & Xiang, 2016; Hofmeister & Norcliffe, 2013). Here we present four experiments which directly test the prediction that resumptive pronouns lead to *more*, but never *less* accurate interpretation than gaps. We do so by measuring how participants interpret sentences with gaps or resumptive pronouns in non-islands, weak islands, and strong islands. If speakers do indeed produce resumptive pronouns when they help comprehenders, then in structures where speakers produce them more frequently, resumptive pronouns should facilitate comprehension more. That is, any facilitation effect should be stronger in islands than in non-islands, but also stronger in strong islands than in weak islands.

In the spirit of doing careful, piecemeal work so as to fully understand the phenomenon, we chose to look at one particular piece of the puzzle: the contribution of parsing to comprehension. We have suggested that Hofmeister and Norcliffe (2013) and Beltrama and Xiang's (2016) pragmatically rich stimuli may have made it possible for their participants to rely on non-compositional strategies for interpretation. In sentences with gaps, participants may have used both parsing and pragmatic cues to interpret sentences, but in sentences with resumptive pronouns, just relying on pragmatic cues may have sufficed to achieve a high rate of accuracy. As a first step at understanding interpretation, then, we designed our stimuli using unfamiliar animal characters (e.g., Miss Rabbit, Mr. Froggy) so as to preclude the use of pragmatic cues during comprehension. Participants therefore had to rely on bottom-up syntactic processing to interpret gaps and resumptive pronouns.

Experiment 1 is a single-trial sentence-picture matching task where participants were presented with a sentence and four images representing possible interpretations. Experiment 2 is a self-paced reading task, a partial replication of Hofmeister and Norcliffe (2013) experiment. Experiment 3 is an eyetracking study using a visual world paradigm, which allowed us to assess online sentence interpretation. Experiment 4 is a single-trial sentence comprehension task. In all experiments, the Facilitation Hypothesis makes the same prediction: resumptive pronouns should make the comprehender at least as likely to correctly interpret sentences as gaps. If, on the other hand, resumptive pronouns result in decreased interpretation accuracy, then they cannot be said to facilitate comprehension. This would be inconsistent with the Facilitation Hypothesis, but consistent with production accounts.

2. Experiment 1: sentence-picture matching

In Experiment 1, a single-trial sentence-picture matching task, we asked participants to select one of four scenes reflecting possible interpretations of a sentence with a gap or a resumptive pronoun, as in Fig. 1. The scenes were all equally (im)plausible, such that reasoning over world knowledge would not help participants to identify the correct interpretation. We used the single-trial method (which has been previously employed and validated; von der Malsburg, Poppels, & Levy, 2018) to avoid inadvertently training participants in these unusual structures (Snyder, 2000).

2.1. Method

2.1.1. Participants

We paid 300 workers from Amazon's Mechanical Turk workforce $0.10 (USD) each for participation. Requirements included that participants learned English before they were 6 years old and that they had not previously participated in the experiment. Subjects were randomly assigned to conditions, such that we collected 50 observations per cell. No participants were excluded.

2.1.2. Factors

Two factors were manipulated, resulting in a fully crossed 2 × 3 design. The first factor, RESUMPTION, had two levels: *gap* or *resumptive pronoun*. The second factor, ISLANDHOOD, had three levels: *non-island*, *weak island*, and *strong island*.

2.1.3. Materials

The single item set is given in Table 1. Each sentence began with an animal character (the head noun; "Mr. Dino") as the head of a relative clause. The sentence ended with a gap ("tickled _") or resumptive pronoun ("tickled him") followed by a prepositional phrase ("with a feather"). The gap/resumptive pronoun appeared in either a non-island, a weak island, or a strong island. Thus, gaps and resumptive pronouns each appeared in environments where participants often hear them and in environments where participants rarely hear them. Each participant read one of the sentences in Table 1, and had to match one of the pictures shown in Fig. 1.

The images reflecting different possible interpretations of the sentences were the same for each participant, although their order was randomized. We coded each image according to the type of interpretation it reflected (codes shown in Fig. 1). The image of the pig tickling the dinosaur is the *target* image, because it reflects an interpretation where the gap or resumptive pronoun refers to its head noun (in this case, the dinosaur). Our best guess about the most likely alternative was that the pronoun would be interpreted as referring to the only other gender- and number-congruent animal in the sentence: the rabbit. We therefore included an image of the pig tickling the rabbit, which we call the *local* interpretation because the rabbit is the closest potential referent. Such an interpretation may reflect a parse favoring local coherence (Tabor, Galantucci, & Richardson, 2004). That is, participants may simply disregard the first few words ("It is Mr. Dino that…") so as to render a clearly grammatical and easy to interpret string ("Mr. Rabbit said that Miss Piggy tickled him with a feather."). The final two images were included to ensure that participants were paying attention and not selecting responses at random. The image of the pig tickling the duck reflected a *dangle* interpretation, where the gap or resumptive pronoun refers to a non-sentential referent (the duck was not mentioned in the sentence). Finally, we called the image of the rabbit tickling the dinosaur the *bonkers* interpretation because there should be no ambiguity as to which character was the subject/agent of the verb "tickle."

2.1.4. Procedure

In all experiments, subjects read instructions, requirements for participation, informed consent, and compensation information prior to beginning the experiment. Experiment 1 instructions stated: "You will be presented with a sentence and four pictures. One picture depicts the scene described in the sentence. Your task is to click on the image that matches the sentence. Participation takes about 1 minute." They then

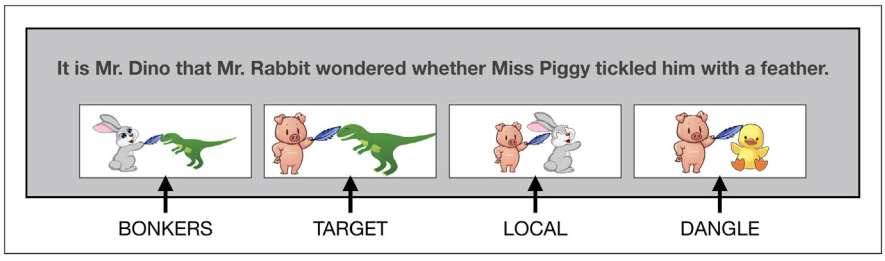

Fig. 1. Sample display from Experiment 1. The trial shown here is a *resumptive pronoun, weak island* condition. Participants were instructed to read the sentence and click the scene which reflected their interpretation of the sentence. The four response options – *target*, *local*, *dangle*, and *bonkers* (labels not shown to participants) – appeared in random order.

Table 1
Experiment 1 stimuli. Sentences appeared in a 2 × 3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.

Islandhood	Stimulus
Non-island	It is Mr. Dino that Mr. Rabbit said that Miss Piggy tickled _/him with a feather.
Weak Island	It is Mr. Dino that Mr. Rabbit wondered whether Miss Piggy tickled _/him with a feather.
Strong Island	It is Mr. Dino that Mr. Rabbit slept while Miss Piggy tickled _/him with a feather.

Note. Because Experiment 1 was a single-item experiment, Table 1 gives all stimuli used in the experiment, not just a representative item set.

followed a link from the Mechanical Turk interface to the experiment, where they saw one of our six stimulus sentences above four images, as in Fig. 1, and clicked on the image that corresponded to their interpretation of the sentence. They had unbounded time to respond. No feedback was given.

2.2. Analysis

Throughout this study, data were analyzed using linear mixed effects models (Baayen, Davidson, & Bates, 2008) with maximal random effects structures. Maximal random effects structures have been shown to prevent inflated false positive effects that can arise with non-maximal, simplified random effects structures, e.g., intercept-only models (Barr, Levy, Scheepers, & Tily, 2013). The downside of maximal models is, though, that they can be difficult to fit in the frequentist framework due to their complexity. One solution, which we use throughout, is to fit linear mixed models in the Bayesian framework.[3] Bayesian linear mixed models, while closely mirroring the conceptual structure of frequentist linear mixed models, have the benefit of gracefully falling back on priors where the analogous frequentist model would fail to converge. Thus, even if there are insufficient data to estimate random effects in the frequentist framework, the Bayesian model can still produce sensible results for the fixed effects of interest.

One consequence of fitting Bayesian linear mixed models as opposed to frequentist mixed models is that the summary statistics take a different form. While a frequentist analysis produces *p*-values quantifying how likely it would be to see the observed effect or something more extreme if the null-hypothesis were true, a Bayesian analysis produces posterior distributions for the parameters in question. These posterior distributions quantify the probability of each possible parameter value in light of the data. From this posterior we calculated three summary statistics which are reported for each analysis.

First, we report the posterior mean $\widehat{\beta}$, the best estimate of the parameter. Second, to give readers a sense of the precision of this estimate, we also report the 95% credible interval ("95%-CrI") – a range around the best estimate $\widehat{\beta}$ that has a 95% chance of including the true parameter value. Third, where a frequentist analysis would use *p*-values to support inferences, we use $P(\beta > 0)$, i.e. the probability that the effect of interest is greater than zero (Nicenboim & Vasishth, 2016). If this probability was above 95%, we concluded that the effect was reliable and positive; if it was below 5% we concluded that the effect was reliable and negative. If it was above 99% (or below 1%), we considered that to indicate *strong evidence* for an effect. If it was between 90% and 95% (or 5% and 10%), we took that as *weak evidence* (at best). Between 10% and 90%, we concluded that there was no evidence for the effect.

In this and subsequent analyses, RESUMPTION was coded using a sum contrast with 0.5 for resumption and −0.5 for gaps. As a result, the parameter estimate for RESUMPTION indicates the expected increase in the dependent variable when a resumptive pronoun is shown instead of a gap. ISLANDHOOD was coded using a treatment contrast with non-islands as the base-level and weak islands and strong islands as treatments (Schad, Vasishth, Hohenstein, & Kliegl, 2020). The parameter estimate for resumption therefore indicates the expected effect in the non-island condition and the interactions indicate how the effect of RESUMPTION differed in weak and strong islands when compared to non-islands. Technical details about model fitting can be found in the Appendix A. All code and data are publicly available on OSF (https://osf.io/9WHN6).

Trials with *dangle* or *bonkers* responses were rare and not of primary interest. We therefore excluded them from all analyses. Thus the dependent variable in Experiment 1 represented whether the response was target (coded as 1) or local (coded as 0), and the data were consequently analyzed with a logistic regression.

2.3. Results

Results of the logistic regression are summarized in Table 2. In non-island conditions, there was evidence that resumptive pronouns significantly decreased *target* responses and increased *local* responses ($\widehat{\beta} = -0.75$, $P(\beta > 0) = .03$; see Fig. 2). There was weak evidence that weak islands elicit fewer *target* interpretations than non-islands irrespective of RESUMPTION ($\widehat{\beta} = -0.44$, $P(\beta > 0) = .09$). Numerically, the effect of resumption was a bit smaller for weak islands, but this interaction was not reliable.

There was strong evidence that strong islands elicit fewer *target* interpretations irrespective of RESUMPTION ($\widehat{\beta} = -1$, $P(\beta > 0) < .01$). Numerically, resumptive pronouns reduced *target* interpretations even more for strong islands than for non-islands, but this difference was not reliable.

2.4. Discussion

Contrary to the prediction of the Facilitation Hypothesis, resumptive pronouns did not lead to more accurate interpretation of sentences than gaps. In fact, they decreased interpretation accuracy and increased locally coherent but globally infelicitous interpretations. This was true in non-island conditions, which is perhaps not surprising given that resumptive pronouns are rarely produced in these contexts. But it was also true in island conditions, where resumptive pronouns are often produced and where prior theoretical and experimental work indicated that resumptive pronouns should have a facilitatory effect.

This hindrance effect, which we will refer to as the *resumptive pronoun penalty*, calls into question the interpretation of Hofmeister and Norcliffe (2013) reading time advantage and Beltrama and Xiang's (2016) subjective comprehensibility rating boost. If resumptive pronouns decrease the likelihood of correct interpretation, then even if resumptive pronouns decrease comprehenders' processing times or increase their confidence in their interpretation, then they do not

[3] In response to the suggestion of an anonymous reviewer, we also report results of frequentist models in supplemental materials. Results of these models pattern with the results we report throughout the paper, although it should be noted that statistical significance in the frequentist models does not map directly to what we report as credible evidence from Bayesian models. This is to be expected given differences between the types of models. One such difference is that the frequentist models had reduced random effects structures to allow convergence. Because of this, we caution the reader that results are not as reliable as the corresponding Bayesian results. We thank Dan Kleinman for contributing helpful code to facilitate convergence of the frequentist models.

Table 2
Experiment 1 results: target vs. local interpreations (see text).

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept (GAP, NONISLAND)	1.5	[0.99,2]	> .99**
RESUMPTION	−0.75	[− 1.6,0.05]	.03*
ISLANDHOOD:WEAK	−0.44	[− 1.1,0.2]	.09
ISLANDHOOD:STRONG	−1	[− 1.7, − 0.39]	< .01**
RESUMPTION × ISLANDHOOD:WEAK	0.12	[− 0.97,1.2]	.59
RESUMPTION × ISLANDHOOD:STRONG	−0.25	[− 1.3,0.81]	.32

Note. Here, $\hat{\beta}$ is the posterior mean, i.e., the best estimate of the effect; 95%-CrI is the 95% percentile credible interval; and $P(\beta > 0)$ is the probability that the true parameter is above zero. $P(\beta > 0) = .03$ means that there is a 3% chance that the true parameter is above zero and a 97% chance it is below zero.
A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

the resumptive pronoun – that is, the *spillover region* of the resumptive pronoun – would be read faster than that of a gap. If participants are at least as accurate in their interpretations of sentences with resumptive pronouns as with gaps, then faster reading times after resumptive pronouns may be evidence in support of the Facilitation Hypothesis. However, if resumptive pronouns lead to fewer correct interpretations relative to gaps, as they did in Experiment 1, we will take this as evidence against the Facilitation Hypothesis.

3.1. Method

3.1.1. Participants

We paid 96 subjects from Amazon's Mechanical Turk workforce $8.00 each for participation. Requirements were that participants learned English and no other language before they were 6 years old and that they had not previously participated in this experiment or

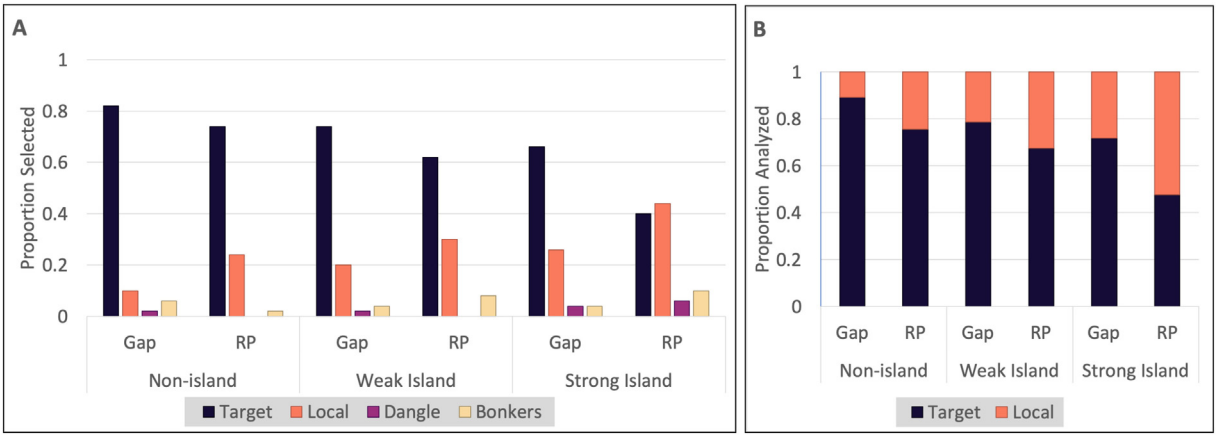

Fig. 2. Experiment 1 results: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis.

facilitate comprehension.

In Experiment 2, we tested whether the resumptive pronoun penalty would extend to a different paradigm. We also aimed to replicate Hofmeister and Norcliffe (2013) reading time advantage for resumptive pronouns, and to ask whether this pattern would hold in island conditions as well as the non-island structures they tested. We also tested whether a within-subjects design might stand a better chance of detecting any processing facilitation associated with resumptive pronouns.

3. Experiment 2: self-paced reading

Experiment 2 was a self-paced reading task. On each trial, participants pressed a button to read sentences word-by-word and then responded to the multiple choice question, “Who did what to whom?” This experiment was designed to be a partial replication both of Experiment 1 and of Hofmeister and Norcliffe (2013) self-paced reading experiment where they found that words after a resumptive pronoun were read faster than words after a gap. Based on Hofmeister and Norcliffe (2013) results, we predicted that the words immediately after

Experiment 1. Five participants were excluded: two because their mean accuracy on unambiguous filler trials was below 40%, one for having participated twice (only the data from the second session were excluded), and two for reporting having learned another language before the age of 6. No exclusions were made on the basis of data collected during critical trials.

3.1.2. Factors

We manipulated the same factors as in Experiment 1: RESUMPTION (*gap* or *resumptive pronoun*) and ISLANDHOOD (*non-island*, *weak island*, *strong island*). This resulted in a fully-crossed 2 × 3 design.

3.1.3. Materials

We created 48 item sets, an example of which is given in Table 3. Aside from our experimental manipulations, every sentence was structurally identical. Each began with a clefted animal character and ended with a gap or resumptive pronoun in direct object position followed by a prepositional phrase introducing an instrument. Characters in the sentence were pseudo-randomly drawn from a pool of eight animal characters such that each character appeared in each argument

Table 3
Experiment 2 stimuli. Sentences appeared in a 2 × 3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.

Islandhood	Sample stimulus
Non-island	It was Miss Piggy that Miss Cat reported that Mr. Dog poked _/her with a pencil.
Weak Island	It was Miss Piggy that Miss Cat understood why Mr. Dog poked _/her with a pencil.
Strong Island	It was Miss Piggy that Miss Cat snacked while Mr. Dog poked _/her with a pencil.

Table 4
Response options for Experiment 2: options listed here correspond to the item set given in Table 3 and were the same for all six conditions.

Label	Sample response options
Target	Mr. Dog poked Miss Piggy with a pencil.
Local	Mr. Dog poked Miss Cat with a pencil.
Dangle	Mr. Dog poked Miss Rabbit with a pencil.
Bonkers	Miss Cat poked Miss Piggy with a pencil.

position a roughly equal number of times across items and such that the filler ("Miss Piggy" in Table 3) and the middle subject ("Miss Cat" in Table 3) were always one gender, while the lowest subject ("Mr. Dog" in Table 3) was always the other gender. Other elements that varied across critical items included the tense of the root clause (half of the critical items began, "It is…," and the other half, "It was…"); the gender of the head noun/resumptive pronoun (half were feminine and half masculine); and the subordinator (half used "that" and half "who"). All logical possible combinations of these features appeared a roughly equal number of times across items. All clause boundaries contained an overt subordinator ("that/why/while" in Table 3).

After every sentence was read word-by-word, the question "Who did what to whom?" appeared with four response options (Table 4). Responses were systematically created to match those of Experiment 1, except these were sentences and not images. The four options always included a *target* interpretation (where the gap or resumptive pronoun refers to the head noun), a *local* interpretation (where the gap or resumptive pronoun refers to the most local gender-agreeing noun, i.e., the middle subject), a *dangle* interpretation (where the gap or resumptive pronoun has an extra-sentential referent), and a *bonkers* interpretation (where the gap or resumptive pronoun is correctly interpreted as the head noun but the subject of the low verb is wrong).

Five sets of twelve filler items (60 total) were also included. Each set was specifically designed to deter subjects from developing a particular type of heuristic parsing or response strategy (e.g., 'the pronoun always refers to the first animal in the sentence'; see supplementary materials for a full explanation of filler types and corresponding results).

3.1.4. Procedure

The experiment was hosted on Ibex Farm (Drummond, 2013). Instructions stated, "In this experiment, you will read about 100 sentences. After each sentence, you will answer a comprehension question. Sentences will be presented to you one word at a time. To go on to the next word, press the spacebar. Please read carefully and do your best to select the correct response. Some sentences will be difficult, so don't worry if you aren't sure. Go with your best guess."

Each trial began with a row of dashes and spaces. Participants pressed the spacebar to reveal the first word, and then pressed the spacebar again to replace the first word with dashes and reveal the next word. After the last word of the sentence, a new screen appeared with a textual comprehension question and four response options. No feedback was given.

3.2. Results

Filler items all included one unambiguously correct interpretation among the four multiple choice options (see supplementary materials). Overall, accuracy on these trials was high: 82% of filler trials were answered correctly (excluding one type[4]), indicating that participants performed the task as intended.

Multiple choice data from critical trials (Fig. 3) were analyzed using logistic regression as in Experiment 1, except that we now included crossed random effects for subjects and items. The overall pattern was similar to Experiment 1. Multiple choice interpretation results are summarized in Table 5. In non-islands, there was weak evidence that resumptive pronouns elicited fewer target responses and more local responses ($\widehat{\beta} = -0.19$, $P(\beta > 0) = .06$). There was also weak evidence that weak islands elicited slightly fewer target responses than non-islands ($\widehat{\beta} = -0.14$, $P(\beta > 0) = .08$), but there was no reliable evidence that the reduction in target responses due to resumption was different from that in non-islands. Strong islands elicited fewer target responses than non-islands ($\widehat{\beta} = -0.58$, $P(\beta > 0) < .01$), and there was no reliable evidence that the reduction in target responses due to resumptive pronouns was different from that in non-islands.

For the reading time data (Fig. 4), we followed Hofmeister and Norcliffe (2013) in defining the critical region as the second word after the gap or resumptive pronoun. In our stimuli, this was always a determiner: the *"a"* in *"with a pencil"* in Table 3). Prior to the reading time analysis, we excluded 9 trials where the critical word was read in more than 5000 ms and an additional 9 trials in which any word in the sentence was read in less than 100 ms. A total of 4590 trials were included in the analysis.

The dependent variable in the reading time analysis was reading speed at the critical word measured in words per second. Thus, the intercept of 3.00 (Table 6) indicates that the word could be read three times per second and that the reading time therefore was 333 ms (Baayen & Milin, 2010; Kliegl, Masson, & Richter, 2010; Wu, Kaiser, & Vasishth, 2018;). The distribution of the residuals was assumed to be Gaussian and this was confirmed using posterior predictive checks.[5]

Results of the reading speed analysis appear in Table 6. Reading speed in the non-island condition (intercept of the model) was estimated to be 3 words per second. There was strong evidence that resumptive pronouns increased reading speed in non-islands ($\widehat{\beta} = 0.28$, $P(\beta > 0) > .99$). Reading speed in weak islands was slower than in non-islands ($\widehat{\beta} = -0.08$, $P(\beta > 0) = .01$), but there was no evidence suggesting that the effect of resumption was different. Similarly, reading speed was overall slower in strong islands than in non-islands ($\widehat{\beta} = -0.08$, $P(\beta > 0) < .01$), and again there was no evidence suggesting that the effect of resumption was different.

3.3. Discussion

The interpretation data from Experiment 2 showed a similar pattern to Experiment 1. The critical effect of resumption – the *hindrance effect* – did not quite meet our threshold for reliable evidence at $P(\beta > 0) = .06$, and is therefore cautiously taken as weak evidence. However, the strongest form of evidence for an effect is consistent replication. As such, the overall similarity of the interpretation data between Experiments 1 and 2 suggest that resumptive pronouns probably did lead readers in Experiment 2 to select fewer target responses and more locally coherent (but incorrect) responses. The similarity between the data from Experiments 1 and 2 also mitigates concerns related to the single-item, between-subjects nature of Experiment 1.

The reading time data in Experiment 2 replicated Hofmeister and Norcliffe (2013) finding that in non-islands, the words after a

[4] One of the five filler types was answered correctly only 32.8% of the time (a conservative estimate of chance would be 25%). These 12 filler items were designed so as to require participants to establish a referent for the pronoun that was not present in the sentence – a *dangle* interpretation. Participants apparently felt a strong pressure to choose a referent for the pronoun from among the characters mentioned in the sentence.

[5] It is common to analyze the logarithm of self-paced reading times. However, examining the residuals usually suggests that this violates the distributional assumptions of Gaussian linear models, and this can in turn distort the results (Kliegl, Masson, & Richter, 2010). We therefore analyzed reciprocal reading times, which has the added benefit of providing parameter estimates that are transparently interpretable in terms of reading speed in words per second.

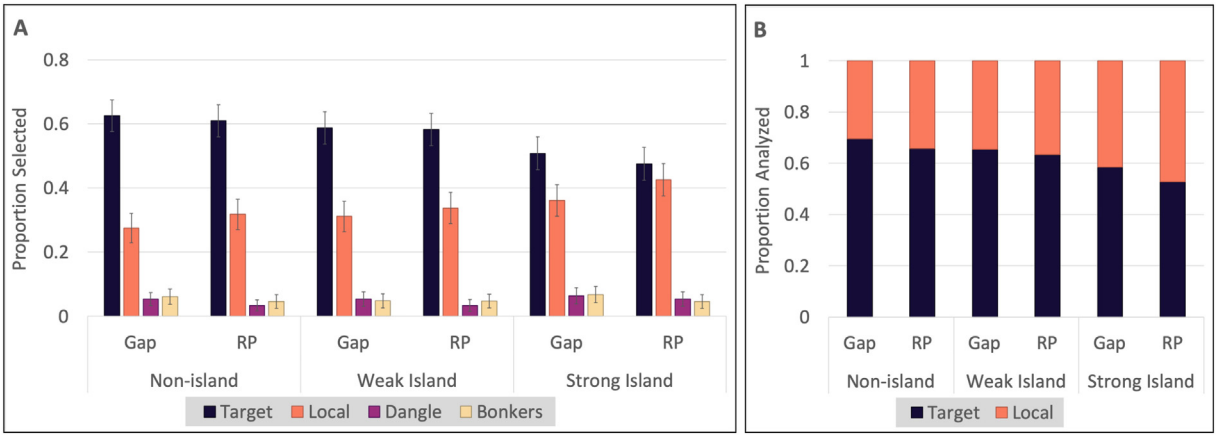

Fig. 3. Experiment 2 results: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis.

Table 5
Experiment 2 results: multiple choice interpretation responses.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept	0.82	[0.59,1]	> .99**
RESUMPTION	−0.19	[− 0.44,0.053]	.06
ISLANDHOOD:WEAK	−0.14	[− 0.34,0.06]	.08
ISLANDHOOD:STRONG	−0.58	[− 0.81, − 0.36]	< .01**
RESUMPTION × ISLANDHOOD:WEAK	0.071	[− 0.29,0.43]	.65
RESUMPTION × ISLANDHOOD:STRONG	−0.089	[− 0.43,0.26]	.3

A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

reading times could not only indicate facilitation, as is most often assumed, but also readers abandoning a parse (e.g., Nicenboim, Logačev, Gattei, & Vasishth, 2016). Indeed, this may be trivially true for any dependent measure, indicating a need for more multi-paradigm studies.

We therefore ran Experiment 3, a visual world experiment where we tracked comprehenders' eyes while they comprehended auditory stimuli. We measured when they looked at which animal characters while they listened to sentences with gaps and resumptive pronouns in order to better understand online processing of resumption.

4. Experiment 3: visual world eyetracking

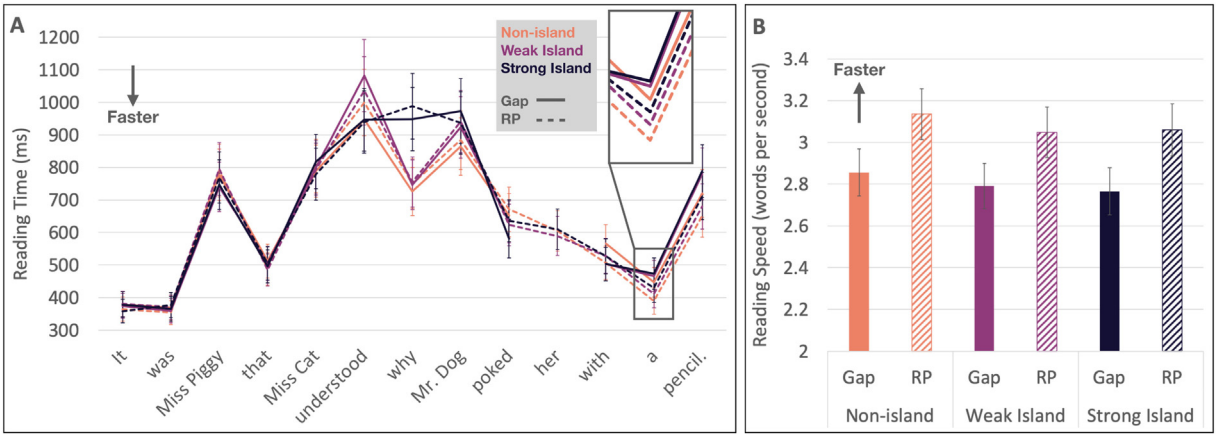

Fig. 4. Experiment 2 self-paced reading results. (A) All reading times (zoom box on the critical region). The word regions are shown with the sample weak island stimulus sentence from Table 3. The corresponding non-island sentence would have “reported” and “that” in place of “understood” and “why.” The corresponding strong island sentence would have “snacked” and “while.” (B) Reading speeds (i.e., the DV in our statistical model) at the critical region.

resumptive pronoun are read faster than the words after a gap. There was no evidence that this effect was different in islands. Taking less time to perform the same process undoubtedly constitutes more efficient processing. But our participants did not perform the same processes when reading gaps and resumptive pronouns. If they had, resumptive pronouns would have shown the same pattern of interpretation as gaps. It is therefore impossible to conclude from these data that resumptive pronouns constitute an improvement relative to gaps from the point of view of the comprehender.

Again, reading times alone do not shed light on the underlying mechanisms: different processes may result in the same pattern of reading times. It is not usually possible to attribute a significant difference to any specific process in the absence of other data: Faster

Experiment 3 was a visual world paradigm. We used the same stimulus sentences as in Experiment 2 (with minor modifications to accommodate the paradigm; see below), but presented sentences auditorily to subjects through headphones while they looked at four animal characters in the corners of a monitor. As in Experiments 1 and 2, we asked participants how they interpreted the sentence at the end of each trial. Response options were identical to those in Experiment 2. If the resumptive pronoun penalty we saw in the previous two experiments is independent of modality (i.e., whether participants read or heard the sentences), then the multiple choice data should again show fewer target interpretations in resumptive pronoun conditions than in gap conditions.

In visual world comprehension studies, the comprehender's gaze

Table 6
Experiment 2 results: reading speed.

	$\widehat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept	3.00	[2.8,3.1]	> .99**
RESUMPTION	0.28	[0.17,0.38]	> .99**
ISLANDHOOD:WEAK	−0.08	[− 0.14, − 0.01]	.01*
ISLANDHOOD:STRONG	−0.08	[− 0.15, − 0.02]	< .01**
RESUMPTION × ISLANDHOOD:WEAK	−0.02	[− 0.15, − 0.02]	.4
RESUMPTION × ISLANDHOOD:STRONG	0.02	[− 0.12,0.15]	.59

Note. A positive coefficient means more button presses per second, i.e. faster reading. For instance a $\widehat{\beta}$ of 1 would mean participants read one additional word per second.
A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

indicates the focus of attention, which in turn is mechanistically driven by comprehension processes. Thus, if we want to know how processing differs when parsing a gap dependency versus a resumptive pronoun dependency, we can compare looks to potential referents while subjects listen to gaps and resumptive pronouns (Altmann, 2004; Altmann & Kamide, 2004, 2009; Altmann & Mirković, 2009; Huettig & Altmann, 2005; Huettig, Rommers, & Meyer, 2011).

To determine whether resumptive pronouns facilitate online comprehension relative to gaps, we compare how accurate referent identification is when processing gaps and resumptive pronouns. If comprehenders' gazes more accurately pick out the target interpretations after resumptive pronouns than after gaps, then this would constitute evidence in support of the Facilitation Hypothesis in online processing.

Given the findings of Experiments 1 and 2, however, we predict that the more likely outcome will be the opposite. We have suggested that comprehenders are simply confused by resumptive pronouns. In self-paced reading data, this would be reflected in decreased reading times as readers try to end the trial more quickly. In visual world data, the confusion account predicts that resumptive pronouns will lead to less accurate referent identification than gaps in the gaze data. Specifically, we predict that comprehenders' looks after hearing a resumptive pronouns will approach chance between the two plausible referents of the pronoun (both the *target* and the *local* referents).

4.1. Method

4.1.1. Participants

We ran subjects from the UC San Diego undergraduate population until we reached our target of 96 participants who met a priori criteria for being included in analyses.

These criteria included: (1) across all trials, participants looked more to the first noun while hearing the first noun than they looked to any other character (1 exclusion); (2) that participants stayed awake for the duration of the experiment (3 exclusions); (3) that participants' responses to multiple choice interpretation questions on unambiguous fillers exceed 80% accuracy (6 exclusions); (4) that each participant provided at least one trial's worth of eye-tracking data during the region of interest (gap/resumptive pronoun and spillover) in each of the 6 cells of the experiment (i.e. the eye-tracker detected their eye and they were not looking at the fixation cross or away from the screen during this portion of every critical item; 2 exclusions); and (5) that the experimental computer and the eye-tracker ran smoothly and without the need for frequent recalibration (as reported by experimenters; 41 exclusions). No exclusions were made on the basis of behaviors contingent on the experimental manipulations.

Subjects received course credit for participation. Pre-screen requirements were that participants were over 18 years old, that they learned English and no other language before they were 7 years old, and that they had normal or corrected-to-normal vision.

4.1.2. Apparatus

An SR Research Ltd. Eyelink 1000 eyetracker was used to record movements of participants' left eyes at a sampling rate of 1000 Hz. The eyetracker was mounted in a tower with forehead and chin rests to constrain head movement. Subjects were seated approximately 60 cm from the screen, on which animal characters appeared in the corners and text (comprehension questions and fixation crosses) appeared in the center.

4.1.3. Factors

We manipulated the same factors as in Experiments 1 and 2: RESUMPTION (*gap* or *resumptive pronoun*) and ISLANDHOOD (*non-island*, *weak island*, *strong island*). This resulted in a fully-crossed 2 × 3 design.

4.1.4. Materials

Critical stimuli for Experiment 3 were the same as those in Experiment 2 except for two modifications. First, instead of using “that” as the first subordinator for half of the stimuli, we exclusively used “who” in Experiment 3 (to facilitate the stimulus recording). Second, so as to control for duration of the auditory stimuli across islandhood conditions, we changed several embedding verbs so that within an item set all embedding verbs had the same number of syllables. Thus, where the sample item for Experiment 2 (Table 3) had “that” immediately after the head noun and contained the embedding verbs, “reported, understood, snacked,” the same item in Experiment 3 (Table 7), had “who” after the head noun and used embedding verbs “reported, understood, exercised,” each of which has three syllables. Response options were identical to those in Experiment 2 (see Table 4).

Visual stimuli consisted of eight digitally drawn animal characters with distinguishing colors, features, and gender-typical clothing and accessories. Auditory stimuli were recorded by a native speaker of American English. Critical stimuli were spliced such that within a given item, all of the lexical content that remained constant across conditions was acoustically identical (i.e., “It was Miss Piggy who Miss Cat,” “Mr. Dog poked,” and “with a pencil” for the item in Table 7). Content that varied across conditions was manipulated using Audacity and Praat so as to have the same durations. For embedding verbs and subordinators (i.e., “reported that,” “understood why,” and “exercised while”), this was achieved by using the *Lengthen* function in Praat to stretch/compress each clip to the mean duration of the original three clips for that item. In cases where this resulted in one or more recordings sounding clearly artificially manipulated, the two-word clips were re-recorded and the process was repeated until recordings were judged by undergraduate RAs to sound like unaltered speech. For gap conditions,

Table 7
Written versions of the auditory Experiment 3 stimuli. Sentences appeared in a 2 × 3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line. Stimuli were almost identical to those in Experiment 2, save for minor changes related to creating controlled auditory recordings.

Islandhood	Sample stimulus
Non-island	It was Miss Piggy who Miss Cat reported that Mr. Dog poked _/her with a pencil.
Weak island	It was Miss Piggy who Miss Cat understood why Mr. Dog poked _/her with a pencil.
Strong Island	It was Miss Piggy who Miss Cat exercised while Mr. Dog poked _/her with a pencil.

silence was spliced in where a resumptive pronoun otherwise appeared such that the duration between the offset of the lowest verb and the lowest preposition was identical in all conditions. To mitigate the oddness of silence in this position, as well as to eliminate coarticulation effects that might make spliced in material sound unnatural, the speaker who recorded the materials produced pauses between words throughout the recording while attempting to approximate normal prosody. The resulting sentences were spoken slowly, with single-syllable words like “who” and “him” averaging 414 ms in duration (including the brief periods of silence mentioned above).[6]

Fillers were identical to those in Experiment 2. In order not to render fillers more or less unnatural sounding than critical items, the recordings were created in a similar way. The speaker produced pauses throughout, and undergraduate RAs used Praat to swap strings with identical lexical content from recordings of other filler items (e.g., “It was Miss Cat who”).

4.1.5. Procedure

We began each experimental session by familiarizing subjects with the animal characters' images and names. Experimenters administered two picture-feature matching quizzes consisting of 16 questions each, for example, “Who wears a yellow bowtie?” and “What color was Miss Duckie's umbrella?” Incorrectly answered questions were repeated at the end of each quiz until all questions had been answered correctly.

Participants put on headphones and the experimenter calibrated the eyetracker. Participants then read the instructions: “In this experiment you will listen to about 100 sentences through headphones while we track your eye movements. We will show you pictures of some of the characters in the sentence. After each sentence, you will answer a comprehension question. To answer comprehension questions, press the number associated with your response. Some sentences will be difficult, so don't worry if you aren't sure. Go with your best guess.” After three practice trials, they began the 108 experimental trials (48 critical, 60 filler), each separated by a brief fixation check. Experimenters monitored calibration for accuracy and recalibrated as necessary.

On each trial, four images of characters appeared in the corners of the screen 500 ms before the onset of the audio recording. On critical trials, the images depicted the three characters in the sentence and the extra-sentential referent of the dangle multiple choice response option (i.e., Miss Rabbit for the item in Table 7). For filler trials, all characters mentioned in the sentence (up to four) were present on the screen; for stimuli with only three characters, a character from a multiple choice response option for that item was selected to appear on the screen as well. The position of characters on the screen was pseudo-randomized such that each interpretation option was equally represented in each corner of the screen across every experimental session (each participant saw the head noun in each corner an equal number of times).

A 200 ms pause occurred at the end of the audio recording after which the four animal characters disappeared from the screen and the question, “Who did what to whom?” appeared in the center of the screen followed by the four response options. Participants pushed the number corresponding to their selection on the keyboard, and then the trial concluded.

4.2. Analysis

Multiple choice responses were analyzed as in Experiment 2 using logistic regression. For the gaze data, we similarly used logistic regression to analyze whether and when participants looked at the target picture as opposed to the local interpretation. For the purposes of this analysis, looks to other regions were disregarded. This was done for two reasons: First, for consistency with the analogous analyses of the multiple choice data in Experiments 1, 2, and 3, and, second, because the research question focused on whether subjects would resolve the gap/resumptive pronoun as being coindexed with either the target (correct) or the local noun (incorrect). Multiple-choice data from Experiments 1, 2, and 3 consistently suggested that these are the options considered by comprehenders. The predictors were RESUMPTION, ISLANDHOOD, and, to account for the possibility that gaze location changed throughout the measurement window, TIME. The measurement window began 200 ms after the offset of the word preceding the gap/resumptive pronoun and ended 1000 ms later. Time was measured in 100 ms steps[7] and the time predictor was then centered and scaled such that −0.5 represented the beginning of the analysis window and 0.5 the end. Parameter estimates for RESUMPTION and ISLANDHOOD therefore indicate differences in the middle of the measurement window at 700 ms post-offset of the word preceding gap/resumptive pronoun. All interactions of RESUMPTION, ISLANDHOOD, and TIME were included in the model. See Appendix A for details.

4.3. Results

Multiple choice data are shown in Fig. 5. Results, summarized in Table 8, were largely the same as in Experiments 1 and 2. There was strong evidence for the resumptive pronoun penalty in non-islands ($\widehat{\beta} = -0.31$, $P(\beta > 0) < .01$). Weak islands elicited about as many target responses as non-islands. Numerically, the resumptive pronoun penalty was reduced (i.e., resumption imposed less of a penalty) in weak islands compared to in non-islands, but there was not enough evidence to conclude that this was a statistically reliable effect. In strong islands, target responses were reduced across the board ($\widehat{\beta} = -0.48$, $P(\beta > 0) < .01$). Interestingly, there was some weak evidence that resumption did not reduce target responses as much in strong islands as in non-islands ($\widehat{\beta} = 0.24$, $P(\beta > 0) = .93$).

The gaze data are shown in Fig. 6 (collapsed across ISLANDHOOD) and Fig. 7 (the gap/resumptive pronoun region broken down in all six conditions). In general, these data were particularly clean. For instance, the left panel of Fig. 6 shows that around 500 ms post-onset of the head noun (“Miss Piggy” in the example sentence), participants' eyes were drawn to the picture of the head noun. The close correspondence between gap and resumptive pronoun conditions indicates that 96 participants was enough to ensure high signal-to-noise ratio. Indeed, one can also see nuanced effects, such as the early drop-off in looks to the gender-incongruent low subject (“Mr. Dog”) after the onset of either of the first two animal characters.

Descriptively, the gaze data from the gap/resumptive pronoun region can be characterized by four observations, labeled in Fig. 6. First, at the onset of the gap/resumptive pronoun, participants' eyes remained on the most recently named character in the sentence (i.e. the low subject, “Mr. Dog” in the example). Second, relative to gaps, resumptive pronouns appear to have resulted in more looks away from the low subject (Mr. Dog). Third, after the onset of a gap, the proportion of looks to the head noun (Miss Piggy/the *target* interpretation) increased, but the proportion of looks to the middle subject (Miss Cat/the *local* interpretation) did not appear to change. Fourth, similar to gaps, after the onset of a resumptive pronoun, the proportion of looks to the head noun (Miss Piggy/*target*) increased, but in contrast to gaps, the proportion of looks to the middle subject (Miss Cat/*local*) also increased and by approximately the same amount. Comprehenders who looked away from the low subject (Mr. Dog) while hearing a resumptive pronoun were more likely to look to the target interpretation than when hearing a gap, but they were also more likely to look to the local interpretation than when hearing the gap. In fact, they appeared to look to the target referent and the local referent with roughly the same

[6] Auditory recordings are available on OSF at https://osf.io/9WHN6.

[7] To confirm that nothing hinged on this choice, we also ran analyses with step sizes of 50 ms and 200 ms and got the same results; also see Appendix.

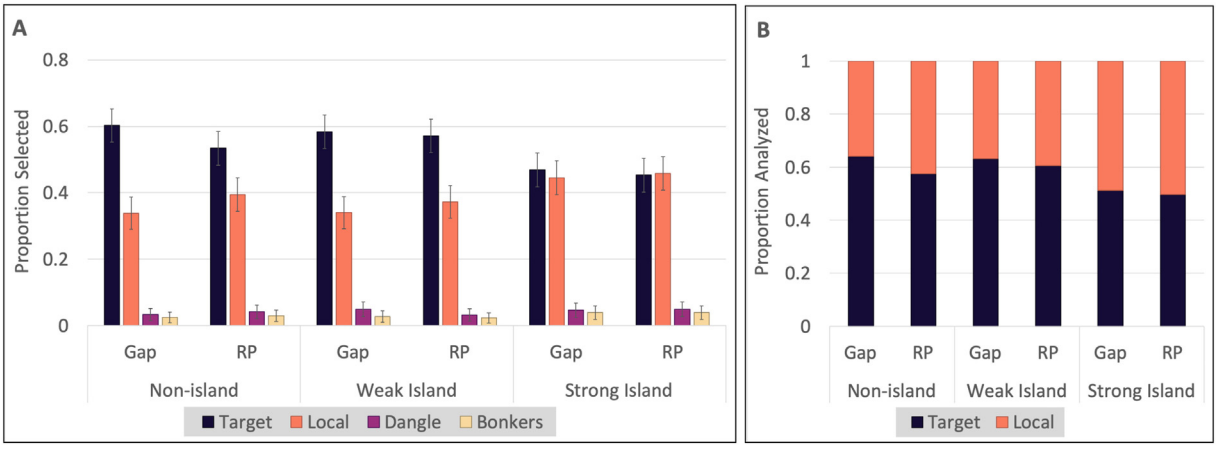

Fig. 5. Experiment 3 results of the multiple choice (interpretation) question: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis.

Table 8
Experiment 3 results: multiple choice interpretation responses.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept	0.49	[0.29,0.69]	> .99**
RESUMPTION	−0.31	[− 0.54, − 0.07]	< .01**
ISLANDHOOD:WEAK	0.07	[− 0.11,0.24]	.76
ISLANDHOOD:STRONG	−0.48	[− 0.66, − 0.3]	< .01**
RESUMPTION × ISLANDHOOD:WEAK	0.17	[− 0.15,0.49]	.85
RESUMPTION × ISLANDHOOD:STRONG	0.24	[− 0.08,0.57]	.93

A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

interpretation data, there was strong evidence that resumptive pronouns reduced looks to the target in non-islands ($\hat{\beta} = -1.1$, $P(\beta > 0) < .01$). Compared to non-islands, weak islands elicited more looks to the target overall ($\hat{\beta} = 0.83$, $P(\beta > 0) = .97$), but there was no evidence suggesting that the resumptive pronoun penalty was different in weak islands than in non-islands. Also similar to what we have observed in multiple choice interpretation data, compared to non-islands, there were fewer looks to the target in strong islands ($\hat{\beta} = -1.9$, $P(\beta > 0) < .01$), and again no evidence that the resumptive pronoun penalty was any different for strong islands than for non-islands. There was also no evidence for an effect of any of the other two-way interactions nor the three-way interaction.

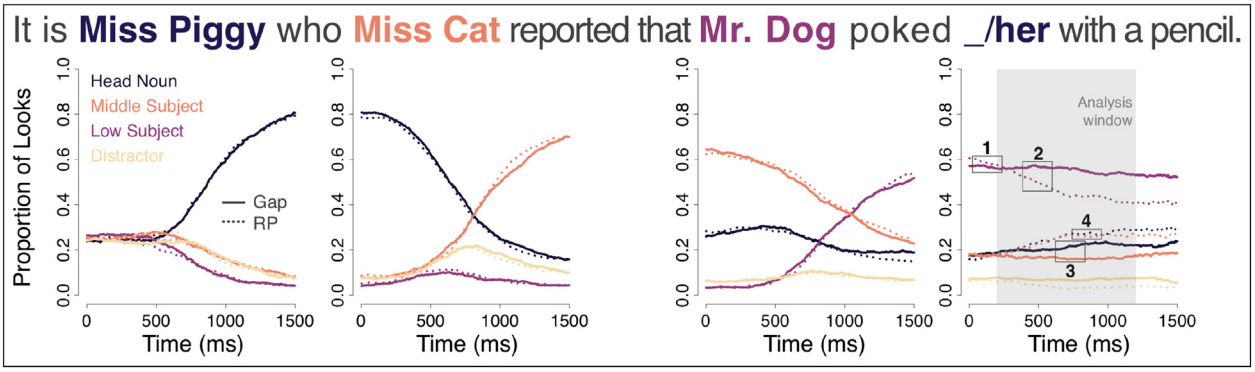

Fig. 6. Experiment 3 gaze data, collapsed across ISLANDHOOD, from the onset of the head noun ("Miss Piggy" in the example sentence), the middle subject ("Miss Cat"), the low subject ("Mr. Dog"), and the gap/resumptive pronoun. The four descriptive points we outline in the text are numbered in the right-most plot: (1) participants' gazes remained on Mr. Dog (i.e. the most recently named character) at the onset of the gap/resumptive pronoun; (2) resumptive pronouns resulted in more looks away from Mr. Dog than gaps; (3) gaps led to more looks to the target than the local interpretation; (4) resumptive pronouns led to roughly equal numbers of looks to the target and local interpretations. The latter two points taken together mean that, although resumptive pronouns led to numerically more target looks than gaps, these looks were less accurate. That is, when participants did look away from Mr. Dog, they were more likely to look at Miss Piggy than Miss Cat when hearing a gap than when hearing a resumptive pronoun.

frequency when hearing a resumptive pronoun, suggesting that resumptive pronouns may be fully ambiguous between these two potential referents in online processing. In other words, resumptive pronouns appear to drive the gaze away from the low subject ("Mr. Dog") but beyond that they don't seem to help comprehenders to identify the correct antecedent.

Results of the statistical eyetracking analysis are presented in Table 9. There was some weak evidence suggesting that looks to the target increased over time in the non-island conditions when collapsing across gap and resumptive pronoun conditions ($\hat{\beta} = 0.28$, $P(\beta > 0) = .92$). Similar to the pattern we have observed in multiple choice

4.4. Discussion

In the multiple choice interpretation task in Experiment 3, resumptive pronouns reduced the number of target interpretations compared to gaps in all three levels of ISLANDHOOD. There was some evidence that this effect was attenuated in strong islands. However, even if this attenuation is real, it was not big enough to counteract the resumptive pronoun penalty, and certainly not big enough evidence facilitation. Resumptive pronouns still led to numerically fewer target responses than gaps in strong islands, meaning that the attenuation does not constitute evidence for facilitation.

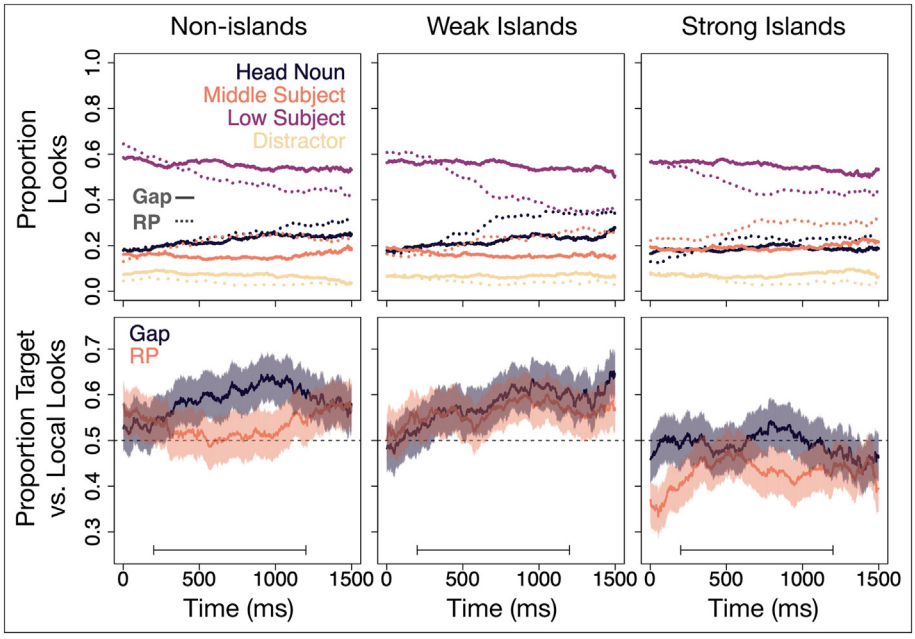

Fig. 7. Gaze data during the gap or resumptive pronoun for all three island types (columns) from Experiment 3. *Top row:* looks to all four characters on the screen. *Bottom row:* looks to the target interpretation, excluding data where participants were not looking at the target or local interpretation (i.e., the dependent variable in our analysis). Shaded area shows standard error. Chance looking between target and local interpretations is 50% (dashed line). Analysis window (200 to 1200 ms) is indicated with horizontal bar in the bottom of each plot.

Table 9
Experiment 3 results: gaze.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept	1.1	[0.27,1.9]	> .99**
TIME	0.28	[− 0.12,0.69]	.92
RESUMPTION	−1.1	[− 2.1, − 0.18]	< .01**
ISLANDHOOD:WEAK	0.83	[− 0.041,1.7]	.97*
ISLANDHOOD:STRONG	−1.9	[− 2.8, − 1.1]	< .01**
TIME × RESUMPTION	−0.24	[− 0.72,0.65]	.47
TIME × ISLANDHOOD:WEAK	0.16	[− 0.42,0.74]	.71
TIME × ISLANDHOOD:STRONG	−0.12	[− 0.67,0.44]	.33
RESUMPTION × ISLANDHOOD:WEAK	0.22	[− 1,1.5]	.63
RESUMPTION × ISLANDHOOD:STRONG	−0.27	[− 1.6,1]	.34
TIME × RESUMPTION × ISL.:WEAK	−0.59	[− 1.5,0.38]	.32
TIME × RESUMPTION × ISL.:STRONG	0.13	[− 0.82,1.1]	.61

A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

This attenuation is another indicator that the effect of resumption is one of increasing confusion such that performance approaches chance. If the increase in local responses for resumptive pronouns reflected a locality preference, then we might expect to see cases in which resumptive pronouns result in more local responses than target responses. Instead, across experiments, the highest rates of local responses for resumptive pronouns (i.e., strong island conditions) are cases where participants select target and local responses at roughly the same rate. This suggests that in these cases, participants are selecting from among the a priori plausible responses (*target* and *local*) at chance.

Overall, the multiple choice data are consistent with Experiments 1 and 2. As the stimuli in this experiment were presented auditorily, we can further conclude that the resumptive pronoun penalty is independent of the modality of stimulus presentation (similar to Clemens, Morgan, Polinsky, & Xiang, 2012 finding that auditory presentation does not change their acceptability).

The gaze data largely replicated the multiple choice interpretation data from this experiment, as well as Experiments 1 and 2. Critically, gaps resulted in more accurate looking behavior than resumptive pronouns. Even though the overall number of looks away from the low subject was less for gaps than resumptive pronouns (see point (2) in Fig. 6), when participants did look away, they looked more to the target referent than the local one when they heard a gap than they did when they heard a resumptive pronoun (points (3) and (4) in Fig. 6). Online as well as offline, then, resumptive pronouns hinder comprehension relative to gaps.

We were particularly surprised by our second descriptive observation (point (2) in Fig. 6) – that resumptive pronouns induced the comprehender to look away from the low subject more than gaps.[8] We speculated that the difference may be attributable to the different ways that comprehenders identify the referents of gaps and pronouns. Pronouns trigger a search for a referent (e.g., Hobbs, 1978; Kaiser, Runner, Sussman, & Tanenhaus, 2009), reflected in looks away from the low subject, which cannot be the referent because it has the wrong gender and is not reflexive (*Principle B* of Government and Binding Theory; Chomsky, 1993). Gaps, on the other hand, can be anticipated because, once a head noun is encountered, the parser can infer that it is in an open wh-dependency. Indeed, Frazier (1987) showed that gaps are actively predicted during sentence comprehension. When the parser finally comes across the gap, the referent of the gap is already known because gaps are syntactically bound to the head noun. No search is needed.

This account straightforwardly predicts our third observation, that gaps lead to more looks to the head noun (*target*) than the middle subject (*local*), because the referent of the gap is a priori known to be the head noun. In resumptive pronoun conditions, on the other hand, participants' eye movements are at chance between landing on plausible referents of a pronoun: the target and local interpretations. This seems to indicate that comprehenders consider both the target and the local interpretations as plausible antecedents, as they would be for an ordinary pronoun. Perhaps this pattern of data arises because resumptive pronouns simply *are* ordinary pronouns from the perspective of the comprehender.

If resumptive pronouns are in fact ordinary pronouns, then the interpretation of ordinary and resumptive pronouns should pattern together, to the exclusion of gaps. Specifically, ordinary pronouns should

[8] A supplemental analysis at 700 ms post-onset showed that resumptive pronouns reliably reduced looks to the lower subject in the non-island condition ($\hat{\beta}$ = −0.29, 95%-CrI: [−0.52, −0.053], $P(\beta > 0) < .01$) and that this effect was even larger in weak islands ($\hat{\beta}$ = −0.39, 95%-CrI: [−0.73, −0.046], $P(\beta < 0) = .01$) and strong islands ($\hat{\beta}$ = −0.34, 95%-CrI: [−0.67, 0.0069], $P(\beta > 0) = .03$).

display the same bias toward local resolution that we have seen for resumptive pronouns. Experiment 4 was designed to test this prediction.

5. Experiment 4: ordinary pronoun comprehension

Experiment 4 used a single-item sentence comprehension task to test the hypothesis that resumptive pronouns are in fact ordinary pronouns from the perspective of the parser. If so, this would be an indicator that there is no grammatical representation of a filler-resumptive pronoun dependency. Resumptive pronouns would be ordinary pronouns from the perspective of the comprehension system.

This would be consistent with production models of English resumption, according to which resumptive pronouns are simply ordinary pronouns from the perspective of the production system (Asudeh, 2004, 2011; Morgan & Wagers, 2018). They are the result of the producer giving up on completing a filler-gap dependency and producing a pronoun where a gap would have otherwise appeared had things not gone awry.

If resumptive pronouns are ordinary pronouns from the perspective of the comprehension system, it predicts that ordinary pronouns will show a similar pattern of interpretation to resumptive pronouns. Specifically, ordinary pronouns should show a local preference for resolution as compared to gaps, as we have observed for resumptive pronouns in the previous three experiments. Participants read a sentence with either a gap, an ordinary pronoun, or a resumptive pronoun and then selected a multiple choice option reflecting their interpretation. Both the sentence and the multiple choice options remained on the screen for the entire trial, as in Experiment 1. Participants were allowed as much time as they needed to respond.

5.1. Method

5.1.1. Participants

We continuously ran workers from Amazon's Mechanical Turk workforce until we reached our target of 150 participants who met a priori criteria for being included in the analysis. A total of 174 participants were run. We excluded 12 for incorrectly answering the comprehension question in the filler trial preceding the critical trial; 10 for reporting that they learned another language before the age of 7; and 2 for responding to either the filler trial or the critical trial in less than 5 s (both of these participants responded in less than 2 s and gave incorrect answers on the filler trial). Participants were paid $0.35 each for participation. Pre-screen requirements included that participants learned English before they were 7 years old and that they had not previously participated in the experiment. Each subject was assigned a different condition from the previous subject, such that we collected 50 observations per cell.

5.1.2. Factors

We included one factor, REFERRING ELEMENT, which had three levels: *gap*, *resumptive pronoun*, and *ordinary pronoun*.

5.1.3. Materials

We created one item set, given in Table 10, with four multiple choice interpretation options, Table 11, which were the same for each of the three stimulus sentences. In order to compare a gap or resumptive pronoun to an ordinary pronoun, we had to make a significant structural change to the sentence. In the ordinary pronoun condition, when the reader reaches the pronoun, the wh-dependency is already resolved; the pronoun must therefore be interpreted as an ordinary pronoun.

In the gap and resumptive pronoun conditions, on the other hand, there must be an unresolved wh-dependency when the reader reaches the gap/pronoun so that the pronoun is interpreted as resumptive. To do this, we introduced a new argument position by using the verb *ask* to embed the lowest clause inside a weak island. Unlike the previous verbs that we used in weak island stimuli (e.g., *wonder whether, understand why, consider whether*) *ask* allows an optional direct object before its clausal complement, as in "ask (someone) whether...." In the ordinary pronoun condition, the dependency terminates in this position with a gap, so the subsequent pronoun is unambiguously an ordinary pronoun.

In the gap and resumptive pronoun conditions, we filled this direct object position of *ask* with the first person pronoun *I*, thereby keeping the dependency open. The choice of *I* was deemed optimal because it cannot be a referent for the resumptive pronoun as the two have different person features (*him* cannot be used to refer to *I*). It also added less of a working memory burden relative to the ordinary pronoun condition than names or full noun phrases would have (Lewis, 1996). Thus, across conditions, the REFERRING ELEMENT had the same syntactic role (direct object), thematic role (patient), and semantic role (the character who was reported to the boss).

Multiple choice options were the same across conditions and appeared in random order. Because the referent of the ordinary pronoun is ambiguous by design, there is no "target" interpretation; we therefore refer to this as the "distant/target" option.

5.1.4. Procedure

The experiment was hosted on Ibex Farm (Drummond, 2013). Instructions stated: "In this experiment, you will answer comprehension questions about 3 sentences. The whole task should take just a minute or two. When doing the experiment we ask that you stay focused and avoid distractions like multitasking. Please do not listen to music with words. Underneath each sentence there will appear four possible interpretations. Select the one that is most likely to be true based on the sentence." Participants started with a practice trial, followed by a filler trial, followed by the critical trial. No feedback was given.

5.2. Results

Data (Fig. 8) were analyzed using a logistic regression to model the

Table 10
Stimuli for Experiment 4.

Islandhood	Stimulus
Gap	It was Mister Bear that I asked Mister Dog why Miss Duckie reported _ to the boss.
Resumptive pronoun	It was Mister Bear that I asked Mister Dog why Miss Duckie reported him to the boss.
Ordinary pronoun	It was Mister Bear that _ asked Mister Dog why Miss Duckie reported him to the boss.

Note. Because Experiment 4 was a single-item experiment, Table 10 gives all stimuli used in the experiment, not just a representative item set.

Table 11
Response options for Experiment 4. As in previous experiments, the comprehension question was, "Who did what to whom?"

Label	Sample response options
Distant/Target	Miss Duckie reported Mister Bear.
Local	Miss Duckie reported Mister Dog.
Dangle	Miss Duckie reported Mister Frog.
Bonkers	Mister Bear reported Miss Duckie.

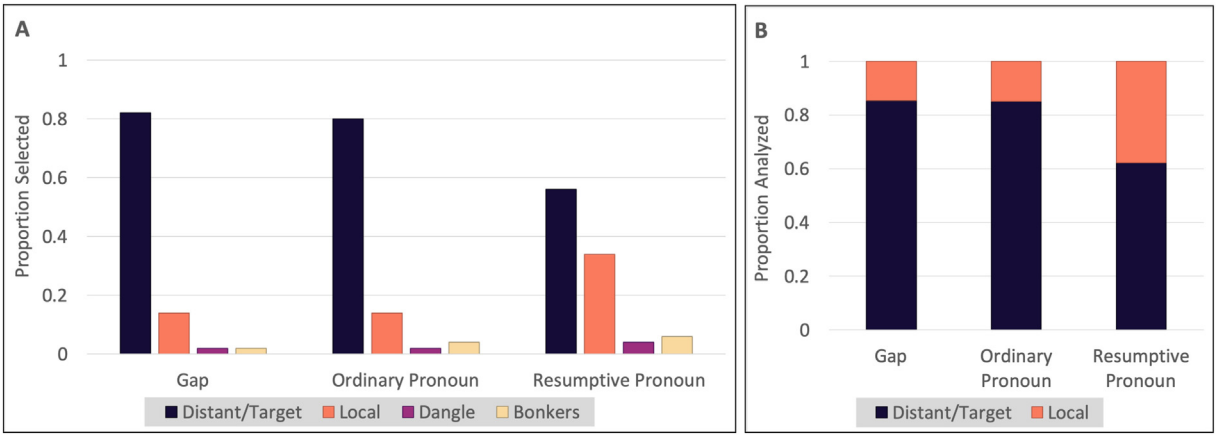

Fig. 8. Experiment 4 results: (A) all responses and (B) just *distant/target* and *local* responses — i.e., those included in the analysis.

Table 12
Experiment 4 results: multiple choice interpretation responses.

	$\widehat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept (ORDINARY PRONOUN)	1.58	[0.95,2.27]	> .99**
GAP	0.14	[− 0.81,1.09]	.61
RESUMPTIVE PRONOUN	−1.01	[− 1.87, − 0.2]	< .01**

A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

rate at which subjects chose the *distant/target* response as opposed to the *local* response. The single predictor was REFERRING ELEMENT (*gap, resumptive pronoun, ordinary pronoun*) coded as a treatment contrast with *ordinary pronoun* as the reference level. Results are summarized in Table 12.

We found no credible evidence that ordinary pronouns are interpreted differently from gaps. We did, however, find evidence that ordinary pronouns are processed differently from resumptive pronouns: ordinary pronouns elicited more distant/target responses than resumptive pronouns ($\widehat{\beta} = -1.01$, $P(\beta > 0) < .01$).

5.3. Discussion

Experiment 4 aimed to assess the hypothesis that English resumptive pronouns are in fact ordinary pronouns and not a kind of alternative gap, as is reported for languages like Hebrew and Irish. Specifically, we tested the prediction that the interpretation data for resumptive pronouns and ordinary pronouns would pattern together, to the exclusion of the gap data, demonstrating a locality bias. Consistent with the first three experiments, resumptive pronouns resulted in decreased distant/target responses relative to gaps and increased local responses. Contrary to our predictions, however, the interpretation of ordinary pronouns patterned with gaps, not resumptive pronouns. Thus, resumptive pronouns appear to involve interpretation processes that are distinct from both gaps and ordinary pronouns.

We believe that the most likely explanation for this pattern is that resumptive pronouns are ungrammatical and confuse the comprehender. When forced to select a referent, comprehenders approach chance in selecting between the gender- and number-congruent discourse entities. Two observations from the data are consistent with this hypothesis. First, we have used the term “locality bias” to refer to the fact that resumptive pronouns result in more local interpretations than gaps. If there were a true bias for local interpretations, however, we might expect to see a condition for which resumptive pronouns lead to more local interpretations than target interpretations. Across four experiments, this pattern never obtained. Instead, resumptive pronouns seem to level out the rates of target and local interpretations, a pattern less consistent with a locality bias than with chance performance, which is indicative of confusion.

Indeed, the effect of islandhood on gap interpretation may serve as proof of concept. Islands induce a similar effect in the way comprehenders interpreted gap conditions. Where gaps are acceptable (i.e., in non-islands), they are understood better than where they are moderately unacceptable (weak islands) and severely unacceptable (strong islands). As target interpretations decrease with decreasing acceptability, local interpretations increase. There is no clear reason that islands might induce local interpretations of gaps; gaps should always unambiguously refer to the filler. A more reasonable interpretation of this pattern is that islands, being ungrammatical, lead to confusion, and comprehenders' performance becomes closer to chance.

A final concern we wished to address regards the generalizability of our findings. We removed pragmatic content from our stimuli to isolate the contribution of parsing. But it also rendered sentences that are subjectively odd and may not behave in the same way as more typical examples of resumption. For instance, it is not clear that English speakers would produce resumptive pronouns in these types of sentences. If resumptive pronouns are not produced in these sentences, then there is no paradox for our stimuli: the (lack of) production and the (lack of) comprehension would in fact be aligned.

Our claim that the Facilitation Hypothesis cannot explain the comprehension-production paradox relies on speakers producing resumptive pronouns in the same kinds of sentences where resumption hinders comprehension. We therefore ran a final experiment to determine whether English speakers produce resumptive pronouns in these sentences.

6. Experiment 5: production

In Experiment 5, a single-trial sentence production task, we asked whether resumptive pronouns are produced in the same types of sentences where we have shown they hinder comprehension. If we are right in claiming that resumptive pronouns are produced as the result of difficulties in production and not because they aid comprehension, then we might expect to see resumptive pronouns produced in these sentences. But if the Facilitation Hypothesis is right, and resumptive pronouns are produced because they facilitate comprehension, then speakers should not produce resumptive pronouns in this experiment because resumptive pronouns do not facilitate comprehension in these stimuli.

Following Morgan and Wagers (2018), we asked participants to type into a text box to complete a sentence that required them to produce a gap or resumptive pronoun. Based on Morgan and Wagers's (2018) findings, we expected that if participants produce resumptive pronouns

Table 13
Experiment 5 stimuli and target responses. Participants were instructed to complete the sentence started by the prompt using the information given by the context sentence.

Islandhood	Stimulus
Non-island	
CONTEXT:	Mr. Rabbit said that Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit said that Miss Piggy...
Weak Island	
CONTEXT:	Mr. Rabbit wondered whether Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit wondered whether Miss Piggy...
Strong Island	
CONTEXT:	Mr. Rabbit slept while Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit slept while Miss Piggy...
Target Responses (all conditions)	
GAP:	tickled _ with a feather
RP:	tickled **him** with a feather

Note. Because Experiment 5 was a single-item experiment, Table 1 gives all stimuli used in the experiment, not just a representative item set.

at all, they should produce very few in non-islands, more in weak islands, and even more in strong islands.

6.1. Method

6.1.1. Participants

We paid 300 workers from Amazon's Mechanical Turk workforce $0.15 (USD) each for participation. Requirements included that participants self-identified as native English speakers and that they had not previously participated in the experiment. Subjects were randomly assigned to conditions, such that we collected 100 observations per cell. A total of 372 participants were run; 72 were excluded for participating more than once (all trials from these participants were excluded). No exclusions were made on the basis of the production task responses.

6.1.2. Factors

We manipulated one factor, ISLANDHOOD, which had three levels: *non-island*, *weak island*, and *strong island*.

6.1.3. Materials

The single item set, given in Table 13, was derived from the Experiment 1 item set (Table 1). Target sentences are identical to the critical sentences in Experiment 1. Context sentences were created by removing the first clause ("It is Mr. Dino that") from the Experiment 1 stimuli and replacing the gap/resumptive pronoun with the head noun, "Mr. Dino." Prompts were created by removing the final verb, the gap/resumptive pronoun, and prepositional phrase from the Experiment 1 stimuli. Target responses were the same for all conditions, given in the bottom panel of Table 13).

6.1.4. Procedure

The experiment was run on Mechanical Turk. Instructions stated: "Help us rephrase some sentences by filling in the blank. Not all sentences will have a clear right or wrong answer. Just do your best!" The instructions and an example trial (not involving gaps or resumptive pronouns) remained on the screen while participants completed the critical trial.

6.2. Data coding and analysis

We coded responses in one of four categories: *gap*, *resumptive pronoun (RP)*, *name*, and *other*. Frequencies and examples of each type of response appear in Table 14.

Because our goal was to determine whether English speakers produce resumptive pronouns in the types of sentences we tested in Experiments 1–4, we were conservative in deciding what types of responses to code as "gap" or "resumptive pronoun." Target responses appear in Table 13. We allowed some minor divergences from these targets which we thought were unlikely to impact the likelihood of producing a resumptive pronoun. These included differences in tense/aspect/mood (e.g., "would tickle," "had tickled," "was tickling") and changes to the prepositional phrase that appeared after the gap/resumptive pronoun (e.g., if it was altered, as in "using a feather," or altogether missing). All other changes were coded as "other" and excluded from analysis. These included changes to the verb ("teased" instead of "tickled") or clause structure (e.g., "with a feather tickled him," "helped tickle him with a feather," "took a feather and tickled him with it").

Only trials coded as *gap* or *resumptive pronoun* were included in the analysis; leaving a total of 221 trials (84 non-islands, 75 weak islands, and 62 strong islands). The analysis differed from previous analyses in that RESUMPTION was our dependent variable, and ISLANDHOOD was the sole predictor. Logistic regression was used for the analysis; ISLANDHOOD was coded as a treatment contrast with *non-island* as baseline (as before).

6.3. Results

Results, shown in Fig. 9 and summarized in Table 15, indicate that resumptive pronouns are indeed produced in these stimuli, and more so in islands than in non-islands ($\widehat{\beta} = 0.74$, $P(\beta > 0) = .98$ for weak islands; $\widehat{\beta} = 1.9$, $P(\beta > 0) > .99$ for strong islands).

6.4. Discussion

Experiment 5 showed that resumptive pronouns are produced in the same sentences where they hinder comprehension. This contradicts the Facilitation Hypothesis, but is consistent with production models and with the idea that resumptive pronouns confuse comprehenders. Our production data are broadly consistent with those of Morgan and Wagers (2018): resumptive pronouns were produced least in the non-island condition, more in the weak island condition, and even more in the strong island condition.

It is worth noting that in the strong island condition, participants produced 56% resumptive pronouns (and 44% gaps). In strong islands in Experiment 1, participants chose the target interpretation of this sentence 72% of the time when it appeared with a gap, but only 48% of the time when it appeared with a resumptive pronoun. Thus, even where resumptive pronouns are produced *more* than gaps, they still hinder comprehension.

7. General discussion

This paper investigated a paradox: English speakers consistently report that resumption is unacceptable, but they nonetheless regularly

Table 14
Experiment 5 coding rubric with frequency of each response type in each of the three conditions and examples.

	Frequency			
Code	Non-island	Weak island	Strong island	Examples
Gap	74	54	27	*had tickled; tickled using a feather*
RP	10	21	35	*tickled him; was tickling him with a feather*
Name	5	6	20	*tickled Mr. Dino; tickled Mr. Dino with a feather*
Other	11	19	18	*used a feather to tickle him; was being tickled by a feather; kept Mr. Dino awake by tickling him with a feather; did some tickling with a feather; lightly teased with a bird's plumage; actually decided to tickle Mr. Dino*

and reliably produce resumptive pronouns. This tension stands to shed a sliver of light into the black box of language.

We investigated two different hypotheses put forth in the literature to explain the paradox. The Facilitation Hypothesis views resumption as the result of speakers trying to be helpful to their listeners by providing an explicit pronoun. Production-based accounts, on the other hand, view resumption as the result of processing gone awry during the production of particularly difficult constructions.

These two explanations make different predictions: If resumption is facilitatory for comprehenders, then resumptive pronouns should be easier to understand than gaps. But if resumption is the result of a mishap during production, then resumptive pronouns should be *harder* to understand.

In four comprehension experiments, we found consistent support for the production-based hypotheses. Specifically, instead of increasing the likelihood that comprehenders land on the target interpretation, resumptive pronouns were more likely than gaps to be interpreted as referring to a local distractor. This was true for offline measures as well as online measures, and even in sentences where resumptive pronouns are produced more often than gaps.

In contrast to all previous studies, in our design we carefully avoided providing pragmatic cues that might have led participants to interpret sentences by reasoning over world knowledge, bypassing the effortful task of parsing these particularly difficult structures (e.g., F. Ferreira, Bailey, & Ferraro, 2002). This allowed us to isolate the contribution of the syntax of resumptive dependencies to the comprehension of these structures. In all our studies, stimuli consisted of sentences describing animal characters interacting in equally (im)plausible ways (e.g., a dog cleaning a duck with a loofa, a cat measuring a dinosaur with a ruler, etc.).

Experiment 1, a sentence-picture matching task, provides the first evidence for the resumptive pronoun penalty. The fact that we found this penalty is especially notable because comprehenders were able to look at the critical sentence and interpretation options simultaneously, and without time pressure for response. Still, accuracy rates were significantly lower for sentences with resumptive pronouns than for those with gaps.

Table 15
Experiment 5 results.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept (NON-ISLAND)	−1.7	[− 2.3, − 1.2]	< .01**
ISLANDHOOD:WEAK	0.74	[0.02,1.5]	.98*
ISLANDHOOD:STRONG	1.9	[1.2,2.6]	> .99**

A single asterisk indicates that P(β > 0) is above .95 or below .05 and therefore meets our criterion for a "reliable" result. Two asterisks indicates that P(β > 0) is above .99 or below .01, which we consider "strong evidence."

The resumptive pronoun penalty was replicated in Experiment 2, when we increased the number of critical trials, presented sentences in a self-paced reading paradigm, and presented multiple choice interpretation options as short sentences instead of pictures. We also replicated Hofmeister and Norcliffe (2013) finding that the words immediately after a resumptive pronoun are read faster than words immediately after a gap. Hofmeister and Norcliffe (2013) interpreted this speed-up as evidence for facilitation. However, at least in our data, faster reading times cannot be taken to reflect facilitation, because participants' interpretations were less correct in resumptive pronoun conditions than in gap conditions.

Experiment 3, a visual world eyetracking study with auditory stimulus presentation, also provided evidence for the resumptive pronoun penalty. Again, the interpretation of resumptive pronouns was worse than gaps, both in online (looks to animal characters) and offline measures (multiple choice interpretation questions).

Experiment 4 suggested that that resumptive pronouns may not be just ordinary pronouns from the perspective of the listener: While the resumptive pronoun resulted in a preference for local interpretation relative to the gap, the ordinary pronoun did not.

Experiments 1–4 consistently showed evidence of a resumptive pronoun penalty, contrary to the prediction of the Facilitation Hypothesis. However, these experiments were not highly naturalistic in

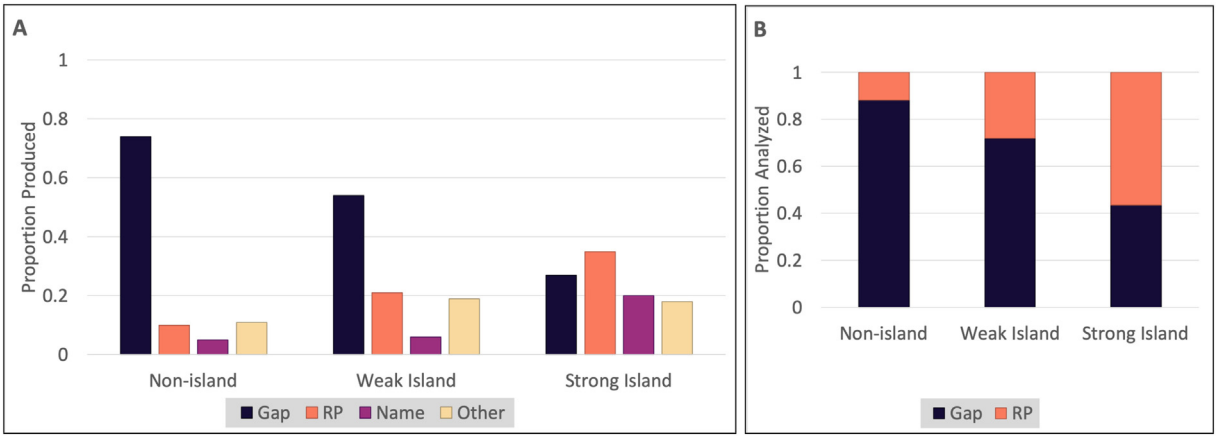

Fig. 9. Experiment 5 results: (A) all responses and (B) just *gap* and *resumptive pronoun* responses — i.e., those included in the analysis.

that sentences were decontextualized and pragmatically impoverished. These may not be the kinds of sentences where resumptive pronouns help comprehension. But if the Facilitation Hypothesis is correct and resumptive pronouns are produced because they help the comprehender, then the strong prediction is that resumptive pronouns should not be produced in these sentences where they do not facilitate comprehension. For this reason, we ran Experiment 5, a production experiment.

Experiment 5 showed that speakers produce resumptive pronouns in the same kinds of sentences where they hinder comprehension. This finding ensures that the resumptive pronoun penalty can be brought to bear on the comprehension-production paradox.

The naturalness of stimuli in comprehension studies is a concern because higher-level logic and reasoning can be brought to bear on comprehension. However, in production, structural choices like whether or not to use a resumptive pronoun are mechanistically guided by lower-level factors like attention and availability of syntactic representations (Levelt, 1993). It is unlikely that these subconscious (V. S. Ferreira, Bock, Wilson, & Cohen, 2008) processes differ from experimental to non-experimental settings. Furthermore the production task was comparatively naturalistic: speakers were given a context sentence and only one trial, so interference from a number of previous trials with repeated use of the same animal characters was not a concern. For these reasons we do not believe ecological validity to be a major concern in Experiment 5.

What these experiments jointly demonstrate is not simply that resumptive pronouns are interpreted with close-to-chance accuracy. This in itself would not be particularly surprising for a structure that is likely ungrammatical. The novel contribution of these experiments is the finding that resumptive pronouns are interpreted less well than gaps, even in islands, where gaps are also ungrammatical, and even in sentences where speakers produce resumptive pronouns. This is exactly the opposite of what was predicted by the Facilitation Hypothesis, leading us to conclude that the Facilitation Hypothesis, at least as generally formulated, is likely wrong. In the following two sections, we expand upon this conclusion, and point to some ways in which the Facilitation Hypothesis may still contain some truth.

7.1. A priori reasons to doubt the Facilitation Hypothesis

As described in the Introduction, previous experimental work has only tangentially tested the Facilitation Hypothesis, and production-based accounts of resumption suggest that resumptive pronouns may be less likely to help comprehenders than gaps. But these are not the only reasons to doubt the Facilitation Hypothesis. Here we spell out three a priori reasons that this account seems unlikely to be correct.

First, a resolution to the acceptability-production paradox must take one of four forms. It can either (i) reject the notion that comprehension and production share grammatical representations (as F. Ferreira & Swets, 2005 do), (ii) do away with acceptability or production (or both) as metrics for grammaticality (see Polinsky, Clemens, Morgan, Xiang, & Heestand, 2013; Shlonsky, 1992 for proposals along these lines), (iii) explain why speakers produce resumptive pronouns in spite of their being ungrammatical, or (iv) explain why comprehenders judge resumptive pronouns to be unacceptable in spite of their being grammatical.

The Facilitation Hypothesis, being a hypothesis about comprehension, does not fit any of these three types. It therefore does not fully address the paradox. The implied logic seems to be: When a speaker senses that an utterance will be confusing or difficult to process with a gap, they opt to produce a resumptive pronoun because, even though it is ungrammatical, the listener is more likely to understand the intended meaning. This would be an explanation along the lines of (iii).

However, research on audience design indicates that speakers do *not* generally take into account the needs of their interlocutors when deciding whether to include optional function words (F. Ferreira & Swets, 2005; V. S. Ferreira & Dell, 2000). For instance, V. S. Ferreira and Dell (2000) show that English speakers produce the optional complementizer *that* not when it would facilitate comprehension on the part of their listener, but when they need more time to plan the following word. Often, the production system seems to be selfish, making decisions to facilitate its own goals, not those of the listener. (See V. S. Ferreira, 2019 for a review.)

Second, it is surprisingly difficult to intentionally produce ungrammatical sentences, and this difficulty is exacerbated if one tries to systematically produce a complex error, for example, one that spans multiple clauses. This anecdote suggests that it is unlikely that such a mechanism could account for the production of resumptive pronouns in English. From a formal perspective, there is no clear way that standard production models can accommodate the production of ungrammatical forms. To put words in the right order to form a relative clause, the system accesses an abstract syntactic representation of relative clauses and then uses the representation to guide lexical selection. But if resumptive pronouns are ungrammatical, then by definition they have no syntactic representation with which to guide production. This is not to say that the production of resumptive pronouns is impossible if they are ungrammatical – Asudeh (2004, 2011) and Morgan and Wagers (2018) provide mechanistic accounts of how this could work. But these accounts do not provide a way for speakers to produce an ungrammatical string by design for the sake of the comprehender, and current general models of production preclude this as a possibility.

Finally, the intuition that resumptive pronouns are more informative than gaps (perhaps due to their overt gender, number, and animacy cues) is probably misleading. English is a language that, for the most part, requires arguments to be pronounced (unlike many other languages, including Spanish, Japanese, and Malayalam). Thus, a missing English noun can usually only be a gap, which must refer to the head noun. Resumptive pronouns, on the other hand, look just like ordinary pronouns, and can therefore plausibly refer to any number of potential referents.

Counterintuitively then, resumptive pronouns are potentially less informative than gaps. Worse yet, if there is no grammatical representation of resumptive pronouns, then there can be no corresponding requirement that the resumptive pronoun refer to the head noun. Whereas comprehenders may infer the intended link between a resumptive pronoun and the head noun, gaps should always lead to more correct interpretation because the grammar explicitly links gaps to the head noun.[9]

Thus, in addition to our experimental findings, we believe there are a number of other reasons that the Facilitation Hypothesis is unlikely to be correct. However, there may still be a sense in which it is correct with some major qualifications. In the following section we discuss this possibility.

7.2. Ways to salvage the Facilitation Hypothesis

Our aim here has been to understand how resumptive pronouns impact comprehension. One feature of our study was that we removed

[9] Note that many theories hold that gaps in islands are ungrammatical, which is to say there is no way to syntactically bind a head noun to a gap in an island. If this is the case, then from the perspective of the parser, gaps and resumptive pronouns are the same: an unbound referential element. (See Phillips, 2006 for a summary of gap processing inside and outside of islands.) Gaps and resumptive pronouns still differ in their gender/number/animacy cues. If this is the only difference, then in islands, resumptive pronouns may in fact be more informative than gaps and may lead to better comprehension. But it is also possible that comprehenders use meta-syntactic knowledge about gaps, which typically may only refer to head nouns. If this is the case, then when there are multiple potential referents with the same gender/number/animacy features, gaps should still be more informative than RPs, regardless of whether they are ungrammatical in islands.

pragmatic information from our stimuli so that any differences in comprehension reflected just the contribution of parsing resumptive pronouns. But it is important to remember that comprehension is not the independent sum of parsing and reasoning. The two may interact in complex ways.

One such way would be discourse context, which we intentionally did not investigate in our studies. For instance, consider how a comprehender might interpret the ordinary pronoun in *"Mr. Bear said that Miss Duckie chased him with a knife."* With a discourse context about a robber, the comprehender might be more likely to interpret the pronoun as referring to the robber (as opposed to Mr. Bear) than if the discourse context had involved, say, a neighbor (depending on one's neighbors; Järvikivi, van Gompel, & Hyönä, 2017; Kaiser, Runner, Sussman, & Tanenhaus, 2009; Koornneef & Sanders, 2013; Williams, Kukona, & Kamide, 2019). Given that neither gaps nor pronouns contain much semantic information and the two do not differ in pragmatic content in any clear way, we think it is unlikely that the use of a gap vs. resumptive pronoun would interact with pragmatic interpretation in such an extreme way. However, given that the syntax of resumption confuses the comprehender, it is possible that comprehenders may rely much more heavily on pragmatic information when processing resumptive pronouns, which could conceivably lead to such an interaction. This is a potentially fruitful avenue for future research.

It is also possible that some other feature of our stimuli has obscured the potential benefits of resumption for comprehension. For instance, it is possible that resumptive pronouns facilitate comprehension only in contexts where they would disambiguate between potential referents. In our stimuli, the only plausible referents for the pronoun were the head noun (*target* interpretation) and the middle subject (*local* interpretation), both of which had the same gender. Resumptive pronouns could therefore not disambiguate between the two. Detailed production data are needed to determine whether such an account holds water: If speakers produce resumptive pronouns more when the pronoun disambiguates, then perhaps the Facilitation Hypothesis is on the right track after all. In this case, our findings would indicate that any facilitation induced by the use of a resumptive pronoun does not come from its impact on parsing, but instead on the comprehender using the number, gender, and animacy cues on the pronoun to resolve the dependency.

Perhaps the most important caveat of all is that if we wish to test the hypothesis that speakers produce resumptive pronouns to help the listener, then comprehension data – be they reading times, comprehensibility ratings, gazes, or responses to interpretation questions – will never fully suffice. Whether or not a resumptive pronoun helps or hinders comprehension is irrelevant when the speaker may have false beliefs about resumption's effect on the comprehender. A true test of the Facilitation Hypothesis will therefore require examining whether speakers are more likely to produce a resumptive pronoun when they believe it will help the listener, regardless of how helpful that pronoun may or may not be in actuality.

7.3. The relationship between comprehension and production

We started by pointing out that resumption leads to an apparent paradox in English: acceptability and production are generally reliable metrics of grammaticality, but they dissociate in the case of resumptive pronouns. Indeed, some early resumption researchers rejected the idea that syntax is shared between production and comprehension. However, two different systems of syntactic representations for production and comprehension would pose a fundamental problem for a variety of behavioral phenomena. There would need to be a – so far unsubstantiated – mapping system linking those two systems, enabling, for instance, structural priming across modalities (Bock, Dell, Chang, & Onishi, 2007; Pickering & Ferreira, 2008; Potter & Lombardi, 1998) and even dialogue (e.g., Pickering & Garrod, 2004).

But whether syntax is shared is not the only issue at stake. A long-standing question in the field is how much else production and comprehension share. In general, models tend to assume that there is a good amount of overlap (e.g., Pickering & Garrod, 2004). As mentioned earlier, Analysis-by-Synthesis theories hold that a large part of comprehension *is* production (Bever & Poeppel, 2010; Halle & Stevens, 1959, 1962). If this is true, then production phenomena like resumption should also appear in comprehension. That is, comprehenders should be able to generate (and therefore interpret) strings with resumptive pronouns.

Thus, if production-based accounts of resumption are correct, then resumptive pronouns pose a challenge to theories that assume comprehension relies on covert production. The paradoxical behavior of resumptive pronouns does not indicate that the two systems do not share grammatical representations, but it may indicate that comprehension and production are more distinct than often thought.

One possible way to rescue an Analysis by Synthesis approach would be if the confusion we see in comprehension resulted from difficulty in mapping between the syntax and semantics of these complex structures. In production, this does not lead to confusion about the message, which is a priori known, but it may explain why speakers produce resumptive pronouns instead of gaps. In comprehension, however, there may be multiple candidate messages that could have prompted the speaker to utter the string of words she did – for example, *Miss Piggy said Mr. Dino tickled Miss Cat* and *Miss Piggy said Mr. Dino tickled Miss Piggy*. The difficulty may not be in covertly producing a structure that matches what the speaker produced, but in maintaining the link between that structure and the specific message that generated it.

Another surprising way in which comprehension and production patterns dissociated in our data is that the resumptive pronoun penalty did not interact with ISLANDHOOD. We predicted that the way comprehenders process resumptive pronouns would be sensitive to syntactic context, reflecting different histories of experience with resumptive pronouns in different structures. For instance, in non-islands, where resumptive pronouns are relatively rare (Morgan & Wagers, 2018), it makes sense that comprehenders would struggle to interpret them. In the comprehender's experience, if a speaker had intended to refer to the head noun in this environment, they would have used a gap, but they didn't, so they must have meant something else.

In weak islands, however, where resumptive pronouns are more common (demonstrated in Experiment 5; see also F. Ferreira & Swets, 2005; Morgan & Wagers, 2018) and rated as more comprehensible (Beltrama & Xiang, 2016), we expected that the resumptive pronoun penalty would be attenuated compared to non-island contexts. Instead, there was no consistent, credible evidence that the resumptive pronoun penalty was different in weak islands from non-islands. The same was true for strong islands.

The fact that comprehenders do not understand resumptive pronouns in islands better again suggests that they are not for the benefit of the comprehender. Producers have access to the intended message, and producing a grammatically licit string has no impact on this. From the perspective of the comprehender, though, when context cues, world knowledge, and pragmatics combine to guide interpretation, they may not always parse complex constructions. When left with nothing but the syntax, they must rely on the parse. Resumptive pronouns, as we have shown, do not provide more helpful information from a syntactic perspective than gaps.

7.4. A new perspective on the competence/performance distinction

The initial perception that the data regarding resumptive pronouns' comprehension versus production are at odds reveals a need for a more nuanced, multilayered framework for understanding performance errors. Based on our results and previous findings, it appears that competence is the same for production and comprehension (i.e., resumptive pronouns are ungrammatical), but that production and comprehension

may have their own performance failure modes that produce characteristic errors.

In the case of resumption, speakers produce resumptive pronouns in response to production pressures. Listeners, on the other hand, must work with what is given: an ungrammatical structure containing an ambiguous pronoun. There is no grammatical parse available, so the listener falls back on other tools. Usually, this includes reasoning over lexical, semantic, prosodic, nonlinguistic and/or contextual cues to resolve referential uncertainty or repair errors in the signal (Levy, 2008; Park & Levy, 2011). But our experiments took most of these tools away: Apart from gender, the listener had no cues to the pronoun's referent, so guesses approached chance between the two gender-congruent characters in the sentence.

In this scenario, there is a performance error in production and we show that comprehenders stumble. But consider two other cases: 1. The *missing VP effect* (Gibson & Thomas, 1999) is the reverse scenario: Comprehenders find double-center embedded RCs with a missing verb (i.e., ungrammatical) just as acceptable as their (grammatical) counterparts without any missing verbs. However, it has not been documented that speakers regularly produce such sentences. So, here the apparent mismatch is a consequence of a performance error just in comprehension. 2. The *depth-charge illusion* is a case where both speaker and comprehender make performance mistakes (Paape, von der Malsburg, & Vasishth, 2019): The speaker produces a sentence like, “No head injury is too trivial to be ignored,” which is compositionally nonsensical. A second performance error in the comprehender creates the illusion that the sentence is well-formed.

To date, cases such as these have all been treated independently. This underscores the field's need for a unified theoretical framework for understanding and explaining performance errors. Ideally, such a framework would account for all the cases mentioned here, and would additionally be able to predict and explain other performance mismatches, which may have yet to be documented.

7.5. Resumption: a cautionary tale for language comprehension research

Finally, our findings invite a methodological point by revealing a weakness in processing studies that measure the time course of language processing without also measuring interpretation (Romoli, Santorio, & Wittenberg, 2020). In the case of resumption, previous researchers had assumed that resumptive pronouns lead the comprehender to the same interpretation as a gap. On this assumption, it may have been reasonable to infer that resumptive pronouns are processed more economically than gaps on the basis of a reading time advantage or increased comprehensibility ratings. At least for the stimuli we tested here, however, faster reading times after resumptive pronouns cannot be interpreted as evidence for processing facilitation. We therefore offer the cautionary guideline for comprehension research: measure interpretation, too.

8. Conclusion

English speakers reliably produce a structure that they deem unacceptable. This is not only odd, but it poses a serious problem for standard assumptions about language and grammaticality. This acceptability-production paradox has spurred considerable curiosity. One prominent hypothesis holds that speakers produce resumptive pronouns because they facilitate comprehension (Ariel, 1999; Asudeh, 2004; Beltrama & Xiang, 2016; Dickey, 1996; Erteschik-Shir, 1992; Fadlon, Morgan, Meltzer-Asscher, & Ferreira, 2019; Hofmeister & Norcliffe, 2013; Prince, 1990).

In four comprehension experiments, however, we show that rather than facilitating comprehension, resumptive pronouns lead listeners and readers to approach chance performance in a variety of interpretation tasks. In a production experiment, we demonstrate that this is true even though speakers produce resumptive pronouns in the same sentences where they hinder comprehension. Our findings contradict the Facilitation Hypothesis, but are consistent with production theories (Asudeh, 2004, 2011; Morgan & Wagers, 2018). While this finding indicates that comprehension and production may indeed share syntactic representations, the paradoxical behavior of resumptive pronouns still imposes a limit on the degree of overlap between the two systems.

We would like to conclude with another puzzle about resumptive pronouns: Languages like Cantonese (Lau, 2016), Hebrew (Ariel, 1999), Irish (McCloskey, 2002), Swedish (Erteschik-Shir, 1992), and Vata (Koopman, 1983) make productive use of gaps, ordinary pronouns, and resumptive pronouns. They do so without the puzzling pattern of acceptability and production data that we have described for English. If resumptive pronouns can be grammatical, then why does English, whose speakers readily produce resumptive pronouns, not simply grammaticize them? That is, if Hebrew speakers produce resumptive pronouns and like the way they sound, then what stops English speakers from doing the same? We suspect that the answer will shed light on something much deeper about human language than just this one quixotic syntactic structure.

CRediT authorship contribution statement

Adam M. Morgan:Conceptualization, Writing - original draft, Visualization, Project administration, Investigation, Data curation, Software.**Titus von der Malsburg**:Writing - review & editing, Formal analysis, Methodology, Conceptualization, Software.**Victor S. Ferreira**:Conceptualization, Methodology, Resources, Writing - review & editing, Supervision.**Eva Wittenberg**:Conceptualization, Methodology, Supervision, Writing - review & editing, Resources, Funding acquisition.

Declaration of competing interest

None. All analyses and data can be found on OSF at https://osf.io/9WHN6.

Acknowledgments

We would like to thank our team of extraordinary RAs for their help in carrying out this research: Suhas Arehalli, Annie Chai, Alexa D'Heilly, Rebekah Hancock-Murphy, Miguel Mejia, Samantha Ngan, Katiana O'dowd, and Talia Orr.

Appendix A. Details of Bayesian data analysis

Bayesian mixed models were fit using the R package *brms* (Bürkner, 2017) which uses the Stan system to obtain posterior distributions (Carpenter et al., 2017). Within Stan, the NUTS sampler (Hoffman & Gelman, 2014) was used to sample from the posterior distributions of the model parameters. We ran four chains each collecting 4000 samples of which the first 1000 were used for warm-up and then discarded. The Gelman-Rubin criterion was used to assess proper mixing of the chains (Gelman & Rubin, 1992).

To avoid overfitting, all parameters received regularizing priors (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; Nicenboim & Vasishth, 2016). For population-level predictors (i.e. fixed effects), we used normal priors with $\mu = 0$ and $\sigma = 1$. Where relevant, group-level parameters (i.e. random effects) were modeled in terms of a correlation matrix and a vector of standard deviations. For the standard deviations we used half-normal priors with $\mu = 0$ and $\sigma = 0.5$. For the correlation matrix an LKJ prior with $\eta = 8$ was used such that smaller correlations among group-level parameters were favored over values closer to the extremes (-1 and 1) without preventing the inference of high correlations if there was evidence for that in the data.

In the analysis of Experiment 2 reading times, the prior for the

intercept was a normal distribution with $\mu = 3$ and $\sigma = 0.5$. This captures the idea that the average reading speed is likely between 2 and 4 words per second. Priors for the other fixed effects were normal distributions with $\mu = 0$ and $\sigma = 0.5$. The prior on the residuals was a normal with $\mu = 0$ and $\sigma = 1$. The other priors were the same as before. This means that if the average reading speed was 3 words per second, most of the data would be between 1 and 5 words per second (i.e. the reading times would be between 200 ms and 1000 ms). Note that these priors are permissive and will be overruled by the data if necessary.

The analysis of eye-movements in visual world experiments is an open problem and various approaches have been used in the literature, from growth-curve analyses, to permutation tests, and logistic regression with polynomial predictors. We loosely follow the general approach described by Barr (2008) first, because it allows us to use simple logistic regression which is familiar to most readers, and second, because this analysis closely resembles the analyses of multiple-choice data in experiments 1 to 4, hence facilitating comparisons across experimental paradigms.

One degree of freedom in this type of analysis is the time interval at which the gaze is sampled.[10] We used 100 ms intervals since this should be enough to capture individual gaze trajectories (fixations are typically considerably longer than 100 ms), but we confirmed that the results also hold with other intervals: When we set the interval to 50 ms or to 200 ms, we obtained the same results. Consult Fig. 7 to confirm that changes in gaze position were low-frequency and 100 ms intervals therefore adequate to capture the overall pattern in the data. Also, given that fixation durations are typically longer than 200 ms, even individual gaze trajectories can be adequately captured with a 100 ms spacing of samples.

Appendix B. Supplementary data

Supplementary data to this article, including information on fillers and the results of frequentist analyses, can be found online at https://doi.org/10.1016/j.cognition.2020.104417.

References

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language,* 110–160.

Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The “blank screen paradigm”. *Cognition, 93*(2), B79–B87.

Altmann, G. T., & Kamide, Y. (2004). Now you see it, now you don’t: Mediating the mapping between language and the visual world. *The interface of language, vision, and action: Eye movements and the visual world,* 347–386.

Altmann, G. T., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. Cognition, 111(1), 55–71.

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609.

Ariel, M. (1999). Cognitive universals and linguistic conventions: The case of resumptive pronouns. *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”, 23*(2), 217–269.

Asudeh, A. (2004). *Resumption as resource management.* Unpublished doctroal dissertationStanford University.

Asudeh, A. (2011). Local grammaticality in syntactic production. *Language From a Cognitive Perspective, 67*, 51–79.

Baayen, R. H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Langauge, 59*(4), 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28.

Barr, D. J. (2008). Analyzing “visual world”eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Beltrama, A., & Xiang, M. (2016). Unacceptable but comprehensible: The facilitation effect of resumptive pronouns unacceptable but comprehensible: The facilitation effect of resumptive pronouns. *Glossa, 1*(1), 1.

Bennett, R. (2009). *English resumptive pronouns and the highest-subject restriction: A corpus study.* UC Santa Cruz: Trilateral (TREND) Linguistics Weekend.

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics, 4*(2–3), 174–200.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition, 104*(3), 437–458.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Cann, R., Kaplan, T., & Kempson, R. (2005). Data at the grammar-pragmatics interface: The case of resumptive pronouns in english. *Lingua, 115*(11), 1551–1577.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures, 9*. Walter de Gruyter.

D’Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology, 19*(5), 381–385.

Clemens, L. E., Morgan, A., Polinsky, M., & Xiang, M. (2012). *Listening to resumptives: An auditory experiment.* New York: Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing.

Davidson Sorkin, A. (2012). Obama wins battleship – with bayonets. Retrieved from https://www.newyorker.com/news/.

Dickey, M. W. (1996). Constraints on the sentence processor and the distribution of resumptive pronouns. *Linguistics in the Laboratory, 19*, 157–192.

Drummond, A. (2013). Ibex farm. Online server *http://spellout.net/ibexfarm*.

Erteschik-Shir, N. (1992). Resumptive pronouns in islands. In *Island constraints* (pp. 89–108). Springer.

Fadlon, J., Morgan, A. M., Meltzer-Asscher, A., & Ferreira, V. S. (2019). It depends: Optionality in the production of filler-gap dependencies. *Journal of Memory and Language, 106*, 40–76.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15.

Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause “island” contexts. *Twenty-first Century Psycholinguistics: Four Cornerstones,* 263–278.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology, 70*, 29–51.

Ferreira, V. S., Bock, K., Wilson, M. P., & Cohen, N. J. (2008). Memory for syntax despite amnesia. *Psychological Science, 19*(9), 940–946.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*(4), 296–340.

Frazier, L. (1987). Syntactic processing: Evidence from dutch. *Natural Language & Linguistic Theory, 5*(4), 519–559.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language & Cognitive Processes, 14*(3), 225–248.

Halle, M., & Stevens, K. (1959). Analysis by synthesis. In *Proceeding of the seminar on speech compression and processing* (Vol. 2, p. D7).

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory, 8*(2), 155–159.

Heestand, D., Xiang, M., & Polinsky, M. (2011). Resumption still does not rescue islands. *Linguistic Inquiry, 42*(1), 138–152.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua, 44*(4), 311–338.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593–1623.

Han, C.-h., Elouazizi, N., Galeano, C., Görgülü, E., Hedberg, N., ... Hinnell, J. Kirby (2012). Processing strategies and resumptive pronouns in english. *Proceedings of the 30th West Coast Conference on Formal Linguistics,* 153–161.

Hofmeister, P., & Norcliffe, E. (2013). Does resumption facilitate sentence comprehension? *The core and the periphery: Data-driven perspectives on syntax inspired by Ivan A* (pp. 225–246). SagCSLI publications.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23–B32.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems Advances in neural information processing systems* (pp. 849–856).

Järvikivi, J., van Gompel, & Hyönä, J. (2017). The interplay of implicit causality, structural heuristics, and anaphor type in ambiguous pronoun resolution. *Journal of psycholinguistic research, 46*(3), 525–550.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition, 112*(1), 55–80.

Keffala, B., & Goodall, G. (2011). *Do resumptive pronouns ever rescue illicit gaps in english.* Poster presented at the 2011 Annual CUNY Conference on Human Sentence Processing.

Kliegl, R., Masson, M. E., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*(5), 655–681.

[10] Gaze position was recorded at 1000 Hz, but the data were thinned out in order to reduce computational cost. As eye movements are relatively slow, keeping samples on a millisecond-basis is not needed.

Koopman, H. (1983). Control from comp and comparative syntax. *The Linguistic Review, 2*(4), 365–391.

Koornneef, A. W., & Sanders, T. J. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language & Cognitive Processes, 28*(8), 1169–1206.

Lau, E. (2016). The role of resumptive pronouns in cantonese relative clause acquisition. *First Language, 36*(4), 355–382.

Levelt, W. J. (1993). Speaking: From intention to articulation (Vol. 1). MIT press.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Stroudsburg, USA: Association for Computational Linguistics.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93–115.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McCloskey, J. (2002). Resumption, successive cyclicity, and the locality of operations. *Derivation and Explanation in the Minimalist Program, 5*, 184–226.

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., & Fedorenko, E. (2018). High local mutual information drives the response in the human language network. *BioRxiv, 436204*.

Morgan, A. M., & Wagers, M. W. (2018). English resumptive pronouns are more common where gaps are less acceptable. *Linguistic Inquiry, 49*(4), 861–876.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology, 7*, 280.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas–part ii. *Lang & Ling Compass, 10*(11), 591–613.

Noble, O. 2017. 24 things Trump does better than anybody (according to Trump). Retrieved from https://news.vice.com/en_us/.

Paape, D., von der Malsburg, T., & Vasishth, S. (2019). Quadruplex negatio invertit? *The on-line processing of depth charge sentences*. Retrieved from https://psyarxiv.com/uw64a under review at journal of semantics. Preprint available at https://psyarxiv.com/uw64a 10.31234/osf.io/uw64a.

Park, Y. A., & Levy, R. (2011). *Automated whole sentence grammar correction using a noisy channel model, 1*, 934–944.

Phillips, C. (2006). The real-time status of island phenomena. *Language,* 795–823.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190.

Polinsky, M., Clemens, L. E., Morgan, A. M., Xiang, M., & Heestand, D. (2013). Resumption in English. *Experimental Syntax and Island Effects, 341*.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language, 38*(3), 265–282.

Prince, E. F. (1990). Syntax and discourse: A look at resumptive pronouns. *Annual Meeting of the Berkeley Linguistics Society, (Vol. 16,*, 482–497.

Ross, J. R. (1967). *Constraints on variables in syntax.* (Massachusetts Institute of Technology).

Romoli, J., Santorio, P., & Wittenberg, E. (2020). *Negation in counterfactuals: What is right and what is not*.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*, 104038. https://doi.org/10.1016/j.jml.2019.104038.

Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic Inquiry, 23*(3), 443–468.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry, 31*(3), 575–582.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*(4), 355–370.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2018). *Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 US and 2017 UK election. PsyArXiv*10.31234/osf.io/n5ywr. Retrieved from psyarxiv.com/n5ywr.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*(8), 989–994.

Williams, G. P., Kukona, A., & Kamide, Y. (2019). (10). Spatial narrative contex8t modulates semantic (but not visual) competition during discourse processing. *Journal of Memory and Language, 108*. https://doi.org/10.1016/j.jml.2019.104030 104030.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*(7), 701–702.

Wu, F., Kaiser, E., & Vasishth, S. (2018). Effects of early cues on the processing of chinese relative clauses: Evidence for experience-based theories. *Cognitive Science, 42*, 1101–1133.

Zaenen, A., Engdahl, E., & Maling, J. M. (1981). Resumptive pronouns can be syntactically bound. *Linguistic Inquiry,* 679–682.

Zukowski, A., & Larsen, J. (2004). *The production of sentences that we fill their gaps. In Poster presented at the 2004 Annual CUNY Conference on Human Sentence Processing.* University of Maryland.