...

Till Poppels* and Andrew Kehler

Ellipsis and the QUD: Sluicing with Nominal Antecedents

https://doi.org/..., Received ...; accepted ...

Abstract: The majority of previous research on ellipsis has broadly agreed that sluicing requires some type of identity (syntactic, semantic, or some combination of both) between the elided material and its antecedent in order to be felicitous. By way of three experiments, we investigate the possibility of inferentially-resolved sluicing, and demonstrate (i) that it can be highly acceptable even when the elided material deviates substantially from its antecedent and (ii) that the acceptability of such sluices varies considerably across items. While we argue that both findings are inherently more consistent with referential theories of ellipsis than identity theories, our investigation also highlights new challenges for referential theories in explaining the variability associated with inferential ellipsis resolution.

Keywords: Ellipsis, Sluicing, Reference, Questions-Under-Discussion

PACS: ...

Communicated by: ...
Dedicated to ...

1 Introduction

Natural languages provide various ways for speakers to minimize their articulatory effort without compromising the expressiveness of their utterance (Zipf, 1949). An extreme form of articulatory reduction is enabled by ELLIPSIS, a set of phenomena that allow a speaker to felicitously produce grammatically fragmentary material when the meaning of the complete thought she wishes to convey can be inferred from the context by the addressee. Unlike other forms of inferential content enrichment (e.g., implicatures), elliptical utterances are those for which the grammar itself provides a tip-off to the hearer that otherwise mandatory material has been elided. Successful communication of

***Corresponding author: Till Poppels,** Université de Paris, LLF, CNRS
Andrew Kehler, Department of Linguistics, UC San Diego

the intended meaning depends on the hearer's ability to recover the missing pieces from the context and combine them with the overt parts of the speaker's utterance.

While there is wide-spread consensus that elliptical utterances depend on the context in *some* way, the nature of this dependency has been subject to decades-long debates among linguists. These debates have been fueled by the fact that the distribution of ellipsis is much more complex than the conceptual simplicity of its communicative purpose—redundancy reduction—may suggest. Explaining how the syntactic, semantic, anaphoric, and information structural properties of elliptical utterances come into contact with the relevant properties of the context and give rise to the complexity of the data remains a looming challenge in linguistic theory.

The approaches that theories of ellipsis have taken on this question can, for the most part, be classified into two broad categories: IDENTITY theories, which posit that ellipsis is only grammatical if the elided material is identical—in some relevant way—to a particular part of the linguistic context known as its "antecedent" (Hankamer & Sag, 1976; Sag, 1976; Chung, Ladusaw, & McCloskey, 1995; Merchant, 2001; Chung, 2006; Rudin, 2019), and REFERENTIAL theories, which maintain that ellipsis sites contain null pro-forms that recover their meaning anaphorically through the same mechanisms that support non-elliptical forms of discourse reference (Hardt, 1993; Kehler, 1993; Jäger, 2001; Barker, 2013; Poppels, 2020).[1] Given this theoretical landscape, a fertile area to examine involves elliptical utterances in which there is a mismatch between the elided material and its antecedent at some level of linguistic representation. For identity theorists, mismatch cases constrain viable definitions of identity by providing potential counterexamples. For advocates of referential theories, they challenge the current state of our understanding of the mechanisms that enable inferential reference resolution.

This paper raises challenges for identity theories by focusing on a particular type of ellipsis known as SLUICING (Ross, 1969), exemplified in (1a)–(1b).

(1) a. Someone wrote a paper about ellipsis and information structure, but I don't remember who (wrote a paper about ellipsis and information structure)
b. Regarding Trump's impeachment, the only question is when (he will be impeached).

Examples like (1a), which have become standard in the literature, satisfy the constraints on identity imposed by all sluicing theories of which we are aware. Alongside such cases,

1 A third class of theories rejects the assumption behind identity theories that the ellipsis site contains unpronounced but otherwise fully represented linguistic material, but doesn't explicitly model the use of ellipsis as a form of discourse reference (Dalrymple, Shieber, & Pereira, 1991; Ginzburg, 1992; Culicover & Jackendoff, 2012; Ginzburg, 2012; Ginzburg & Miller, 2019). To what extent those theories are compatible with the assumptions behind the referential accounts of ellipsis cited above is an open question that is beyond the scope of this chapter.

however, we consider novel examples such as (1b), which lie outside of the constraints imposed by identity theories as well as various hybrid proposals (e.g., Chung, 2006, 2013). By way of three experiments, we show that cases like (1b) can be rated as highly acceptable despite the fact that they require inference to recover their interpretations. The results also reveal considerable variability across items, raising new challenges for identity and referential accounts alike. We discuss the ramifications of these data for both types of analyses, and outline directions for future work.

2 Background

The large majority of previous research on sluicing has broadly agreed that some type of identity constraint between the elided material (in brackets in (1a)–(1b)) and the antecedent material (underlined) is required for the felicitous use of sluicing.[2] However, there is less agreement on what the precise nature of this constraint is, with some identity theorists defining it over syntactic representations (Sag, 1976; Chung et al., 1995; Rudin, 2019), some in purely semantic terms (Dalrymple et al., 1991; Merchant, 2001; Potsdam, 2007), and yet others with reference to both syntactic and semantic levels of representation (Chung, 2006; Tanaka, 2011a, 2011b; Chung, 2013; Merchant, 2013b; AnderBois, 2014).

Historically, data involving syntactic mismatch between the antecedent and putative elided material have been central to such debates in the ellipsis literature, since they provide a diagnostic for the level of representation at which identity constraints hold. Specifically, such mismatches should render sluicing unacceptable only if it is (at least partially) sensitive to the syntactic representation of the antecedent, but not if it is governed by a purely semantic identity condition (Lipták, 2015, inter alia). Merchant (2001) in fact famously argued against the purely syntactic identity condition imposed by the analysis of Chung et al. (1995), drawing in part on cases that involve lexical mismatches but preserve semantic identity, such as the one between *the Thompsons* and *them* in (2).

(2) Susan hates the Thompsons and the detectives wanted to know whether [the Thompsons]$_1$ know why (Susan hates them$_1$).

(adapted from Merchant, 1999, ex. 16)

2 We use standard bracketing notation to indicate the optionality of linguistic material as a function of ellipsis, and "#(…)" to mark material that cannot be felicitously elided.

Sluicing is felicitous in (2) despite the fact that Principle C of Binding Theory prevents the definite description *the Thompsons* from reappearing at the ellipsis site and instead requires a pronominal reference (Chomsky, 1986). The result is a lexical mismatch between two co-referential, yet syntactically distinct, forms—a type of mismatch termed "vehicle change" by Fiengo and May (1994). Such cases motivated Merchant to propose a purely semantic identity criterion based on his notion of e-GIVENness, which holds in examples like (2) and thus correctly predicts that ellipsis should be felicitous despite the lexical mismatch.

However, purely semantic identity conditions such as e-GIVENness are not sufficient for capturing the distribution of sluicing either. For example, as pointed out by Chung (2006), whereas sluicing is perfectly felicitous in (3a), the version in which the ellipsis clause lacks the required preposition – as in (3b) – is infelicitous.

(3) a. Joe was murdered, but we don't know by who (Joe was murdered).
(Chung, 2006, ex. 18b)

b. Joe was murdered, but we don't know who(m) #(Joe was murdered by).
(Chung, 2006, ex. 19b)

This effect is unexpected under purely semantic identity theories such as e-GIVENness, since the existence of a murder entails the existence of a murderer, and hence the requisite entailment relations required by the constraint are satisfied (Merchant, 2013b).

An influential solution to this problem, introduced by Chung (2006) and later adopted in several subsequent works (e.g., Merchant, 2008; Tanaka, 2011a; Chung, 2013; Merchant, 2013a, 2013b), is to extend semantic identity by a lexico-syntactic condition known as the No New Words constraint.[3] As the name suggests, this condition prevents the ellipsis of any lexical material that is not provided by the antecedent, which captures (3b) as well as other cases involving stranded prepositions with no correlate in the antecedent:[4]

(4) a. Jackson is jealous, but we don't know who #(he is jealous <u>of</u>).

3 Rudin (2019) proposes an analysis in which the syntactic identity requirement is restricted to the highest elided *vP*—what he calls the EVENTIVE CORE—which allows for the possibility of mismatches in modality, polarity, tense, finiteness, and the like. This proposal also derives the No New Words constraint as well as some form of semantic identity, even though he doesn't explicitly reference either.

4 Whereas Chung's constraint rules out (3b) due to the need to have the "New Word" *by* appear in the ellipsis clause when both clauses are passive (ia), one might ask how the possibility in which the antecedent is passive and the ellipsis clause is active (ib) is ruled out, since in this case it appears that no new words are required:

(i) a. Joe was murdered, but we don't know who(m) #(Joe was murdered <u>by</u>).
b. Joe was murdered, but we don't know who #(<u>murdered</u> Joe).

b. Jackson is jealous, but we don't know of who (he is jealous).
c. Susan said she was afraid, but she didn't say what #(she was afraid of).
d. Susan said she was afraid, but she didn't say of what (she was afraid).

In order to prevent the No New Words constraint from being overly restrictive, Chung (2006) had to carve out several exceptions that are also echoed in Rudin (2019), illustrated in (5).

(5) a. John is eating, but I can't see what$_i$ (he is eating t_i).
b. Susan hates the Thompsons and the detectives wanted to know whether [the Thompsons]$_1$ know why (Susan hates them$_1$). (=(2))
c. [Which person]$_1$ will win the next election and by what margin (will they$_1$ win it)? (Ginzburg, 1992, ex. 302a)

First, the No New Words condition must ignore syntactic traces in cases of sprouting—a subclass of sluicing in which the sluiced *wh*-phrase has no overt correlate in the antecedent (Merchant, 2013a)—such as the trace in (5a). Second, cases involving "vehicle change" (Fiengo & May, 1994) between pronouns and their antecedents, as in (5b), have to be explicitly exempted from the No New Words constraint as well.[5] Finally, pronouns must also be exempted from the No New Words constraints in cases like (5c), even though they are not co-referential with any element in the antecedent. This exception differs from the first two in that these lexical mismatches not only violate the No New Words constraint, but also e-GIVENness: since the elided pronoun receives a non-interrogative interpretation but its correlate in the antecedent receives an interrogative one, the two cannot be semantically identical (Rudin, 2019).

In this paper, we will consider additional challenges to hybrid identity theories of ellipsis that invoke a lexico-syntactic condition like the No New Words constraint. Like (5c), many of the cases we will examine violate both the No New Words constraint as well as semantic identity, but unlike "vehicle change" mismatches, the cases we will consider do not lend themselves to generalized exceptions. Consider the examples in (6), which are representative of the larger set of sluices that feature in the experiments reported in this paper.

Chung's answer is to follow Hale and Keyser (1993) in assuming that the active and passive forms of verbs are different lexical entries, which implies that the active form of *murdered* in (ib) violates the No New Words constraint.

5 This is particularly significant considering the fact that, as mentioned above, Merchant's e-GIVENness was explicitly motivated by the desire to avoid such an exception: "To pursue a theory of [syntactic identity] while considering the cases of 'vehicle change' to have been sufficiently dealt with simply by naming them is to confuse the diagnosis with the cure" (Merchant, 2001, p. 25). Thus, the return to a (partially) syntactic identity condition undermines one of the key motivations for e-GIVENness.

(6) a. A: Can I borrow your textbook?
B: Which textbook (do you have in mind; do you want/need to borrow; . . .)?
b. A: Did you not tell your friends about the game today?
B: I did, but I forgot to tell them where (it would be; it would take place; . . .).

In (6a), B responds to A's request to borrow their textbook with the sluiced clarification question *Which textbook?*, which is readily interpreted as *Which textbook do you have in mind?* or *Which textbook do you need?*. If the ellipsis was licensed under identity with the antecedent clause, however, the sluice would have to be interpreted as *Which textbook can you borrow?*, which is clearly not the interpretation it receives. As we will see, experimental participants interpret these and a variety of other sluices like them in a way that goes beyond the antecedent-provided meaning—i.e., deriving their interpretations under identity would violate both e-GIVENness as well as the No New Words constraint—and yet they nonetheless rate these utterances as highly acceptable. In (6b), the antecedent is a single noun phrase—*game*—rather than a full clause. Since the remnant wh-phrase in the ellipsis clause must compose with a proposition in order to yield a full question meaning, however, the ellipsis must be resolved to a full proposition, not just the meaning of a noun phrase. Therefore, the interpretation of such nominal-antecedent sluices will inevitably have to be inferential to some extent; in this case, the sluiced question may end up meaning something like *where the game would be* or *where the game would take place*.

Our investigation will proceed as follows. Experiment 1 will seek to verify the intuitive judgments about preposition-stranding cases that motivated the No New Words constraint, exemplified in (4), and compare these examples to a novel set of sluices that are modeled after (6a) and are prone to violating both e-GIVENness and lexico-syntactic identity. We ask participants to both judge and paraphrase the ellipsis clause, which allows us to estimate the number of elided lexical items that are not provided by the antecedent and thus must be exempted from the No New Words constraint in order to satisfy identity. Experiments 2 and 3 will then zero in on a subset of the examples in Experiment 1 that involve nominal antecedents. In particular, we will test whether the predictability of the to-be-sluiced questions affects the acceptability of sluicing—a hypothesis that is echoed in several recent theories of sluicing.

3 Experiment 1: sluices with inferred readings

The purpose of this experiment is twofold: to experimentally verify the armchair judgments reported in the literature that motivated the No New Words constraint, and to test it against a novel set of sentences that may receive interpretations that involve New Words and nonetheless appear to be relatively acceptable. We examine short dialogues and employ an experimental paradigm that both elicits acceptability judgments and—in the second part of the experiment—asks participants to paraphrase the ellipsis site in order to determine the meaning it acquires.

3.1 Methods

3.1.1 Materials

We constructed a set of 50 experimental items: 10 minimal pairs that we refer to as CLASSIC items because they were modeled after sentences from the literature that motivated the No New Words constraint, 20 novel items that we refer to as INFERENCE items because they were designed to facilitate inferential ellipsis resolution, and 20 filler sentences that involved sluicing but that did not incentivize inferential readings or involve preposition stranding. Half of the fillers were acceptable while the other half were unacceptable due to a variety of grammatical violations.

All experimental and filler items consisted of a 2-turn dialogue in which the first utterance provided the antecedent and the second utterance contained a sluiced question. The CLASSIC items consisted of minimal pairs like (7), in which the ellipsis clause was paired with an antecedent clause that either did or didn't include a correlate to the elided preposition. The INFERENCE items were designed with the goal of eliciting inferential readings, as exemplified in (8). Finally, examples of the upper and lower bound fillers are shown in (9), and the full list of all experimental items can be found in Appendix A.

(7) a. Claire: Joe is jealous of someone.
Jessica: Do we know who? [acceptable CLASSIC item]

b. Claire: Joe is jealous.
Jessica: Do we know who? [unacceptable CLASSIC item]

(8) Fan: Can I get a few autographs?
Manager: Sure, how many? [INFERENCE item]

(9) a. A: You didn't answer my question.
B: Which question? [upper bound FILLER item]

b. Spectator 1: It's possible that they will score a field goal from there.
Spectator 2: Maybe, but I'm not sure whether. [lower bound FILLER item]

To be clear, the INFERENCE items form a heterogeneous set of examples that were chosen based on our intuitions that they are acceptable under interpretations that would require new words if they were linguistically represented at the ellipsis site. Our goal is therefore not to uncover a new pattern that generalizes to all instances of sluicing. Instead, it is to see whether the No New Words constraint generalizes to a larger set of examples beyond the more limited set of cases cited in the literature.

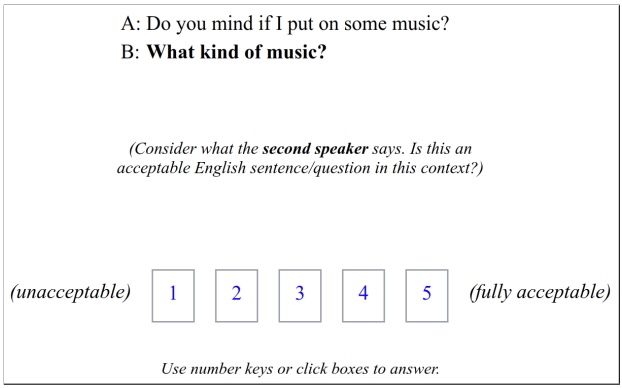
3.1.2 Procedure

Participants performed two tasks in separate blocks. First, they were instructed to evaluate the acceptability of the elliptical utterance ("what the second speaker says") in the context of the first utterance on a scale from 1 ("unacceptable") through 5 ("fully acceptable"). In the second block, participants were asked to paraphrase the utterances they had previously seen in the context of the acceptability judgment task. Specifically, they were asked "What exactly does the second speaker mean?" and told: "While you may use your own words as you type your response, please try to express as precisely as possible what the speaker meant to say." The instructions for both tasks were also included in abbreviated form on every trial to ensure that participants knew what they were asked to do throughout the experiment. The screenshots in Figure 1 show an example trial involving each task from the perspective of the participants.

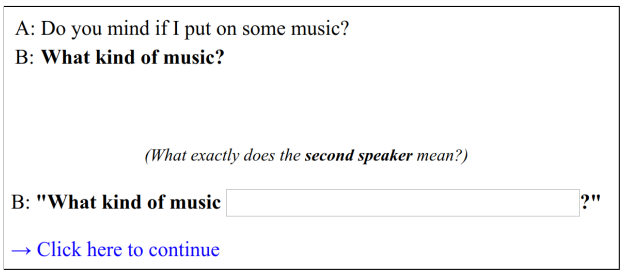
3.1.3 Analyses & predictions

The acceptability results were analyzed by fitting a hierarchical ("mixed-effects") ordinal regression model using the `brms` R package (Bürkner, 2017, 2018). Since we are interested in the differences between the three types of items that were included in the experiment (ACCEPTABLE CLASSIC, UNACCEPTABLE CLASSIC, and INFERENCE items), we included item type as a treatment-coded population-level predictor. As usual, we also added crossed by-subject and by-item group-level intercepts as well as slopes for each population-level predictor. We predicted the novel set of INFERENCE items to pattern with the acceptable variants of the CLASSIC items rather than their unacceptable counterparts.

The second dependent variable, the number of New Words contained in each paraphrase, was calculated by comparing the paraphrase to the antecedent clause and counting how many of its words did not occur anywhere in the antecedent clause. This



(a) Acceptability rating task



(b) Paraphrase task

Fig. 1: Screenshots of the acceptability rating task (top) and the paraphrase task (bottom).

was done in a way that ignored obvious spelling errors (e.g. “unload” vs. “unloadt”), differences in upper/lower case, morphological differences (e.g. “say” vs. “said”), contractions (“he’s” vs. “he was”), and other differences in spelling with no bearing on the No New Words constraint (e.g. “interesting-looking” vs. “interesting looking”). We further removed 71 responses from participants who used the text box to paraphrase the entire ellipsis clause, including the remnants, rather than merely filling in the “missing” material. Finally, we never counted “do it” or “do that” as New Words. Once the number of New Words was determined for each paraphrase, we analyzed these measures by fitting a hierarchical poisson regression model with maximal random effects using `brms` (Bürkner, 2017, 2018) described in more detail below.

We expected the CLASSIC items to be bimodally distributed in terms of both acceptability and the number of New Words: by design, the acceptable variant of each minimal pair includes the target preposition in the antecedent clause, which should lead to fewer New Words in participants’ paraphrases. Since the Number of New Words measure is a novel measure, this prediction will serve to verify the task and

establish a baseline against which the other types of items can be compared. As for the INFERENCE items, we expect a relatively large number of New Words if they do, in fact, succeed in eliciting inferential interpretations. If the No New Words constraint is the correct explanation for the unacceptability of the lower-bound CLASSIC items, and if the INFERENCE items do involve New Words, we should find them to be unacceptable as well.

3.2 Results

3.2.1 Data exclusion

We recruited 28 participants via Amazon.com's Mechanical Turk. One of those participants was excluded for self-identifying as a non-native speaker of English. The data from the remaining 27 participants was further filtered to exclude 9 individual trials on which the response was submitted in less than 1000 ms (we assume that reading and judging the materials carefully would take at least 1 second), leaving us with a total of 1341 individual observations.

3.2.2 Number of "New Words"

We calculated the number of words in participants' paraphrases that did not occur in the antecedent clause following the procedure described above. Figure 2 shows the average number of words per item. We analyzed the results using a multi-level poisson regression with item type (ACCEPTABLE CLASSIC, UNACCEPTABLE CLASSIC, and INFERENCE items) as a treatment-coded population-level effect and all by-item and by-participant group-level ("random") effects justified by the design (Barr, Levy, Scheepers, & Tily, 2013). As expected, the acceptable variants of the CLASSIC minimal pairs involved significantly fewer New Words than their unacceptable counterparts ($\Delta = -0.86$, $CI(\Delta) = [-1.2, -0.51]$, $P(\Delta > 0) = 1$; average difference: 0.86 words). This is an important result because it suggests that the paraphrase task is capable of revealing the difference we are interested in, even though the absolute numbers may not be interpretable directly: while even paraphrases of the acceptable variants contained on average 1.07 words not provided by the antecedent, their unacceptable counterparts involved on average an additional 0.86 such words. That being said, our key prediction was that the INFERENCE items would involve a large number of New Words, and that is exactly what was found: items in this group involved on average 2.67 New Words, significantly more than both the acceptable CLASSIC items ($\Delta = 0.73$, $CI(\Delta) = [0.47, 1]$,

$P(\Delta > 0) = 1$) as well as their unacceptable counterparts ($\Delta = 1.59$, $CI(\Delta) = [1.32, 1.86]$, $P(\Delta > 0) = 1$).

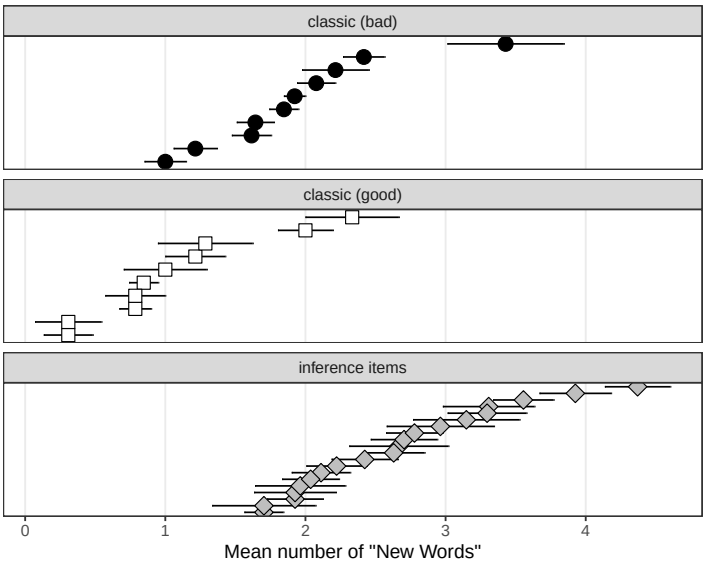


Fig. 2: Mean number of New Words in participants' paraphrases of the ellipsis site, grouped by item type: good and bad variants of CLASSIC minimal pairs from the literature (top and middle); and INFERENCE items (bottom). Errorbars represent Standard Errors.

3.2.3 Acceptability

Figure 3 shows the average acceptability of experimental items across categories. The CLASSIC items (top two facets) exhibit the expected bimodal distribution. The INFERENCE items (bottom) pattern with the acceptable CLASSIC items, even though paraphrasing their meaning resulted in a relatively large number of New Words. We modeled the results in a multi-level ordered-category model with item type as a treatment-coded population-level effects and the maximal random-effect structure, consisting of by-item and by-subject group-level intercepts and slopes for item type (Barr et al., 2013). As expected, ACCEPTABLE CLASSIC items were significantly more acceptable than UNACCEPTABLE CLASSIC items ($\Delta = 2.25$, $CI(\Delta) = [1.71, 2.81]$, $P(\Delta > 0) = 1$), and so were the novel inference items ($\Delta = 1.99$, $CI(\Delta) = [1.21, 2.77]$, $P(\Delta > 0) = 1$). Finally, inference items did not differ significantly from the ACCEPTABLE CLASSIC ones ($\Delta = -0.26$,

$CI(\Delta) = [-0.99, 0.43]$, $P(\Delta < 0) = 0.77$), which is consistent with our expectation that they would pattern together in terms of acceptability.

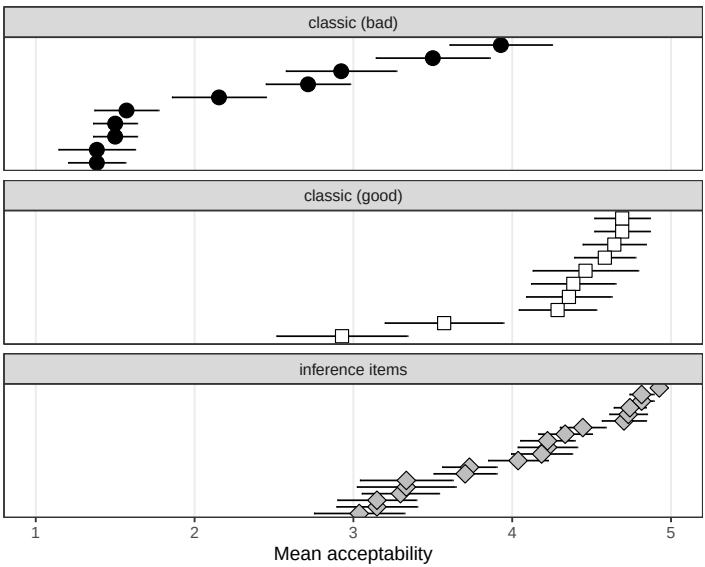


Fig. 3: Acceptability results from Expt 1, grouped by item type: good and bad variants of CLASSIC No New Words minimal pairs from the literature (top and middle); and novel INFERENCE items (bottom). Errorbars represent Standard Errors.

3.2.4 Acceptability as a function of the number of New Words

The scatter plot in Figure 4 shows the relationship between the mean number of New Words and the average acceptability associated with each item. The negative slopes on the lines that connect the two variants of the CLASSIC minimal pairs indicate that with respect to those items, an increase in the number of New Words contained in paraphrases of the ellipsis was associated with a dramatic reduction in acceptability. However, this plot also reveals that this relationship does not hold with respect to the INFERENCE items: they involve an even greater number of New Words but were nonetheless rated as relatively acceptable.

To quantify these observations statistically, we conducted a multi-level analysis with raw ratings as the dependent variable and the number of new words as a continuous predictor along with a 2-way ITEM TYPE factor distinguishing the classic items from the

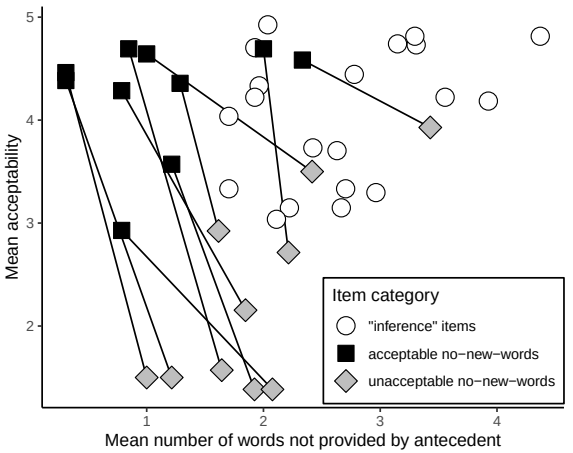


Fig. 4: Scatter plot showing results from Experiment 1. The mean acceptability of each item (y) is plotted in terms of the average number of New Words it produced in the paraphrase task (x). Lines connect variants of the same CLASSIC minimal pair.

inference items and the interaction between them. As always, we added all group-level intercepts and slopes for items and participants that were justified by the design. The results confirm the pattern shown in Figure 4: within the CLASSIC items, increasing numbers of New Words were associated with a significant drop in acceptability ($\Delta = -0.65$, $CI(\Delta) = [-0.84, -0.48]$, $P(\Delta < 0) = 1$), whereas the opposite was true for INFERENCE items ($\Delta = 0.72$, $CI(\Delta) = [0.52, 0.95]$, $P(\Delta > 0) = 1$).

3.3 Discussion

The goal of Experiment 1 was to test the No New Words generalization against (a) preposition-stranding examples sourced from the literature, and (b) a novel set of items that were specifically designed to encourage comprehenders to draw inferences. The CLASSIC items behaved exactly as the No New Words generalization would predict: within each minimal pair, the variant that led to an increased number of New Words in participants' paraphrases of the ellipsis site was also the variant that was associated with a dramatic decrease in acceptability. However, the generalization failed to predict the behavior of the novel set of INFERENCE items that were specifically designed to facilitate inferential ellipsis resolution. We therefore conclude that whatever is causing CLASSIC items to become unacceptable, it cannot be a general ban against eliding words that are not provided by the antecedent.

In fact, the possibility of inferential ellipsis resolution represents a challenge for identity theories of sluicing more broadly: all of the INFERENCE items in this experiment violate identity in one way or another and should consequently be ungrammatical. The fact that most of them are highly acceptable, then, is inconsistent with the identity requirement.

The results from Experiment 1 raise an important question: What supports the inferential resolution of sluicing? In the following section, we take a closer look at a subset of the INFERENCE items from Experiment 1 and develop a working hypothesis that is rooted in referential theories of sluicing and the availability of a salient Question Under Discussion (QUD). The remainder of this paper will be concerned with testing this hypothesis across several experiments.

4 Sluicing and question predictability

4.1 Sluicing with nominal antecedents

One of the most highly rated examples in Experiment 1 is shown in (10), along with the interpretation most participants adopted (in brackets).

(10) A: Did you not tell your friends about the game today?
B: I did, but I forgot to tell them where (the game is).

The overwhelming majority of participants in Experiment 1 interpreted the sluiced question as "where the game is/was." That interpretation, and the fact that the sluice was rated as highly acceptable, would be entirely unsurprising if the context contained an antecedent clause that denotes this proposition, as in (11).

(11) A: Did you not tell your friends that the game is today?
B: I did, but I forgot to tell them where (the game is).

But unlike (11), the context in (10) does not provide a clause that corresponds directly to the meaning of the ellipsis site. Instead the noun phrase *game* appears to serve as the antecedent for the ellipsis clause, which then simply goes on to ask a *where* question about the game.

The fact that a mere NP can serve as the antecedent for clausal ellipsis raises a serious challenge for identity theories of sluicing since noun phrases and clauses are fundamentally distinct linguistic objects, both semantically and syntactically. Despite their theoretical significance, however, nominal-antecedent sluices have received almost no attention in the literature (but cf. Beecher (2007)), which is particularly striking given the fact that analogous facts have received ample attention in the literature on VP-ellipsis

(Hardt, 1993; Kehler, 1993; Johnson, 2001; Merchant, 2013a; Miller & Hemforth, 2014). Consider examples (12a)–(12c).

(12) a. The letter deserves a response, but before you do (respond), ...
(Gregory Ward, p.c., cited in Kehler, 1993, ex. 20)

b. Meanwhile, they sense a drop in visitors to the city. Those who do (visit the city), they say, are not taking cabs.
(Gregory Ward, p.c., cited in Kehler, 1993, ex. 21)

c. Harry used to be a great speaker, but he can't (speak) anymore, because he has lost his voice. (Hardt, 1993, ex. 114)

Each of these examples is felicitous despite the fact that the elided VP and its nominal antecedent are distinct both syntactically and semantically. A common response from identity theorists has been to argue that each of these nominal antecedents underlyingly contains a VP that serves as an identity-preserving antecedent (Fu, Roeper, & Borer, 2001; Johnson, 2001; Merchant, 2013a). Whereas this line of reasoning is unsatisfying with respect to VP-ellipsis, it is even more problematic when applied to sluicing: Whereas NPs like *response*, *visitors* and *speaker* can be reasonably analyzed as deverbal, i.e. as deriving from their corresponding verbs, the NP *game* is not "de-clausal" in the same way, and there are no independent reasons for analyzing it as underlyingly containing a full clause to serve as the antecedent for the sluice.[6]

To make matters even more complicated, any analytic strategy that categorically permits nominal-antecedent sluices like (10) is challenged by the fact that the same context that enables the *where* sluice does not appear to be adequate for sluicing a question with a different remnant wh-phrase:

(13) A: Did you not tell your friends about the game today?
B: I did, but I forgot to tell them how long #(the game is).

We are thus in a position where the ellipsis of the same clause (*the game is t*) in one and the same context with the same nominal antecedent is felicitous for some remnants, but infelicitous for others.

One question that is immediately raised by this contrast is whether the judgments may be driven, or at least modulated, by the plausibility of the sluiced question given its context. As we will see in the next section, this hypothesis is consistent with a line of theories that link the use of ellipsis to the availability of a relevant QUD (Ginzburg & Sag, 2000; Roberts, 2012) in the context of utterance.

6 See Hardt, Anand, and McCloskey (2020) for a proposal that nonetheless pursues this type of strategy.

4.2 Sluicing and the Question Under Discussion (QUD)

QUD-based analyses of ellipsis are not limited to sluicing. In fact, a number of authors have posited that various patterns associated with VP-ellipsis can be analyzed in terms of the extent to which the context raises a relevant QUD (Kertz, 2013; Miller & Pullum, 2013; Miller & Hemforth, 2014; Kehler, 2016, inter alia). Most relevant for our purposes is Miller and Hemforth's (2014) work on VP-ellipsis with nominal antecedents. Whereas the mismatch between the elided material and its antecedent in such cases violates the identity condition (both semantic and syntactic variants), Miller and Hemforth (2014) propose that nominal-antecedent VP-ellipsis instead depends on the extent to which the antecedent NP raises a concealed question, as indicated in square brackets in examples (14a)–(14c).

(14) a. Mubarak's survival [= whether he will survive] is impossible to predict and, even if he does (survive), his plan to make his son his heir apparent is now in serious jeopardy. (Miller & Hemforth, 2014, ex. 1)

b. The integrity of the Senate depends on her participation [= whether she participates]. If she does (participate), . . .
(Miller & Hemforth, 2014, ex. 10a)

c. The release of this information on the user's part depends on his consent [= whether he consents]. If he does (consent), . . .
(Miller & Hemforth, 2014, ex. 10c)

According to Miller and Hemforth (2014), VP-ellipsis in these examples is enabled by the fact that the processing of the antecedent NP requires comprehenders to access the meaning of the concealed question, which makes it accessible for subsequent ellipsis.[7]

7 A reviewer rightly remarks that examples (14a)–(14c), while readily interpretable, might not be quite as impeccable as counterparts with syntactically-matching antecedents (e.g, of the form shown in brackets). As we have pointed out elsewhere (Kehler & Ward, 2007; Kehler, 2018, 2019a; Poppels, 2020), this is precisely the behavior one expects when interpretation is mediated by referential processes. Consider (i)–(ii).

(i) a. Jean is from France, but he hasn't been there in years. (Kehler, 2018, ex. 29a)
b. Jean is French, but he hasn't been there in years.
(Ward, Sproat, & McKoon, 1991, ex.31b)

(ii) a. Do the reactions of parents affect their children?
b. Do parental reactions affect their children? (Ward et al., 1991, appendix, ex. 26)

Whereas reference using *there* in (ib) is clearly acceptable, it is perhaps not quite as acceptable as (ia), where the referent is directly introduced. A similar situation holds for the adjectival antecedent *parental*

Importantly, however, it is not the antecedent NP itself that enables VP-ellipsis, as demonstrated by the following minimal pair:

(15) a. The letter deserves a response, but before you do (respond)...
(Kehler, 1993)

b. The letter contains a response, but before you do #(respond)...

Whereas the context in (15a) succeeds in raising the QUD that is to be addressed by the ellipsis clause, (15b) does not, even though it contains the same antecedent NP *response*. It thus appears that the antecedent alone does not determine whether or not ellipsis is felicitous. Rather, it must conspire with other information provided by the context in order to raise the relevant QUD.

QUD-based approaches to sluicing have also appealed to a notion of QUD accessibility (Ginzburg & Sag, 2000; AnderBois, 2010, 2014; Barros, 2014). For example, AnderBois's (2014) influential theory of sluicing posits a QUD-based wellformedness condition according to which sluicing is felicitous only if the sluiced question corresponds to a salient QUD. In order for this sort of theory to avoid circularity and hence have explanatory value, there has to be an independently motivated measure of QUD availability that is grounded outside the theory of ellipsis. AnderBois addresses this issue by drawing on concepts from Inquisitive Semantics: he credits the inner antecedent with the primary responsibility for raising the QUD that enables sluicing. For example, the inquisitive semantic contribution of the (wide-scope) indefinite *something* in (16) is responsible for raising the QUD *What did Janine hear?* and thus allows that question to be sluiced.

(16) Janine heard something, but she doesn't want to say what (she heard).

While this approach has the advantage of defining precisely how QUDs are raised, it is overly restrictive in a way that undermines the generality of the proposal. For example, it is widely assumed that QUDs can be raised in a variety of ways and not just through the presence of inquisitive linguistic elements like indefinites and disjunctions (Roberts, 2012). Indeed, AnderBois' approach is insufficient for addressing examples of the sort that are of interest here, for instance the contrast between (10) and (13), repeated in (17).

---

in (iib) as compared to (iia). There can be no doubt that the morphological form of the antecedent affects the accessibility of the underlining concept being evoked for subsequent reference in these cases, and no theorist that we are aware of would use these facts to argue that the interpretation of pronouns like *there* in (i) and *their* in (ii) are syntactically-mediated in the way that many syntacticians claim that sluicing is. The same logic applies to any degradation witnessed in (14a)–(14c) compared to counterparts that contain syntactically-matching antecedents: whereas such judgments suggest that morphology impacts the salience of underlying concepts for future reference, they do not provide evidence of an interpretation process that engages with syntax directly.

(17) A: Did you not tell your friends about your game today?
B: I did, but forgot to tell them...
a. ...where (the game is).
b. ...how long #(the game is).

Since there is no inner antecedent, neither the *where* question nor the *how long* question correspond to a salient QUD under his account, making it impossible to explain the contrast between them.[8]

Instead of relying on the presence of inquisitive elements, we will take a different approach towards defining QUD availability. Specifically, we will not only conduct an acceptability judgment study, but pair it with a separate experiment in which participants are presented with the same prompts and perform a forced-choice passage completion task. This will allow us to estimate comprehenders' beliefs about the likelihood of each of the sluiced questions independently of sluicing, which we can then correlate with the acceptability of sluicing those questions.[9]

5 Experiment 2a: nominal-antecedent sluices

The goal of this experiment is to test the acceptability of sluices with nominal antecedents, which will lay the groundwork for testing whether the acceptability of such sluices depends on the availability of the relevant QUD. To that end, we constructed a set of experimental items in a way that allowed us to hold the context of utterance constant while varying the sluiced question and how predictable that question is given the context. In Experiment 2b, we will use the same contexts to estimate how predictable each question is and test whether that predictability score is correlated with the acceptability of sluicing.

8 AnderBois (2014) does suggest that cases without inquisitive inner antecedents may instead rely on bridging inferences, but he does not offer a theory of bridging that would drive a wedge between the examples we are considering here.

9 This strategy is similar to Miller and Hemforth's (2014) approach to nominal-antecedent VP-ellipsis: they experimentally measured the extent to which the antecedent NP gave rise to a concealed polar question and correlated this score with the acceptability of using VP-ellipsis.

5.1 Methods

5.1.1 Materials

We constructed 30 sluices that were similar to example (10) in that they all involved a nominal antecedent. In particular, we constructed six contexts and substituted in five different sluicing remnants, as shown in (18). The complete list of items can be found in Appendix B.

(18) When you call 911 about an emergency, the first thing they ask is...
   a. ...where.
   b. ...why.
   c. ...what exactly.
   d. ...who.
   e. ...when.

The goal behind this item writing procedure was to expose variability in acceptability: while (18a) seems reasonably acceptable, (18b-c) may be somewhat degraded and (18d-e) even more so. Note, however, that the entire context, including the (nominal) antecedent as well as the sluice-embedding clause, was held constant; only the sluicing remnants were varied across item variants.

5.1.2 Participants & procedure

We recruited 63 participants via Amazon.com's Mechanical Turk, all of whom self-identified as native speakers of English. Each participant saw six of the 30 experimental items (one from each context) along with 12 filler items, half of which were designed to be acceptable (19a) while the other half was meant to establish the lower end of the acceptability scale (19b).

(19) a. Mr. Henderson ate either a hamburger or a hotdog, but I don't know which one. [upper bound FILLER item]
   b. Erica married a Russian man and I can't imagine how rich. [lower bound FILLER item]

5.2 Results

We excluded data from trials that lasted less than 1,000 ms under the assumption that it would take at least that much time to read each sentence and submit an acceptability

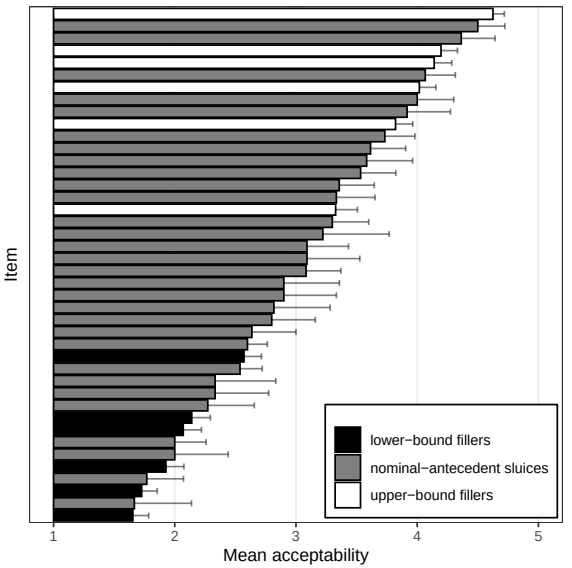


Fig. 5: Results from Experiment 2a. Bars show mean by-item acceptability ratings of acceptable fillers (white), unacceptable fillers (black), and nominal-antecedent sluices (grey). Error bars indicate Standard Errors.

judgment, leaving us with a total of 1012 individual observations to analyze. The results, which are summarized in Figure 5, reveal a large amount of variation: some nominal-antecedent sluices were rated to be highly acceptable, some were judged highly unacceptable, and the remaining items covered the entire range of possible acceptability ratings in-between.

5.3 Discussion

We designed the stimuli for this experiment with the explicit goal of exposing variability that we can then try to explain in terms of QUD predictability. The results reveal that the item creation procedure was successful in this regard: the experimental items covered the entire range of possible acceptability judgments between the bounds established by the filler items.

These results speak to and represent a challenge for all existing theories of sluicing. For identity theories, they are problematic because all nominal-antecedent sluices are predicted to be categorically ungrammatical, whereas some were found here to be highly acceptable. Furthermore, these results cannot be explained by "fine-tuning" the definition

of identity: if we allow sluices with nominal antecedents in order to explain the fact that some of them are highly acceptable, we face the opposite challenge of explaining why others are completely unacceptable. In fact, the extreme gradience that is evident in the results calls into question whether binary classification can be successful at all.

The results are also informative for referential theories of sluicing. According to such theories, contexts with nominal antecedents can give rise to felicitous sluicing as long as the intended referent is sufficiently salient in, or is otherwise readily inferrable from, the context. Since these theoretical constructs are themselves gradient, the variability in the results could at least potentially be explained under referential theories. The goal of the next experiment is to attempt to do so in terms of QUD predictability.

6 Experiment 2b: QUD predictability

6.1 Introduction

The purpose of Experiment 2b is to directly test the hypothesis that the acceptability of nominal-antecedent sluices can in part be explained as a function of the predictability of the sluiced question. To that end, we need to quantify the question predictability associated with each item, and we will do so in a forced-choice passage-completion experiment.

6.2 Methods

6.2.1 Materials, participants, and procedure

The experimental materials were identical to those in Experiment 2a: 30 nominal-antecedent sluices, consisting of six minimal sets with five variants each that differed only with respect to the sluicing remnants. 54 participants were recruited via Amazon.com's Mechanical Turk and presented with six of the 30 nominal-antecedent sluices, one from each context. Unlike Experiment 1, however, the items were truncated before the remnant wh-phrase, followed by 5 possible continuations presented in a forced-choice format, as illustrated in (20).

(20) When you call 911 about an emergency, the first thing they ask is...
   a. ...where you are located.
   b. ...why you are calling.
   c. ...what exactly the emergency is.
   d. ...who they are speaking to.

e. ...when the emergency happened.

Each continuation expressed the meaning of one of the five different sluiced questions that occurred together with this context in Experiment 2a, and were determined in a separate norming experiment described below. Participants were instructed to choose the most likely continuation in this context.

6.2.2 Norming experiment: estimating the meaning of the sluiced questions

In order to estimate the meaning of each of the sluiced questions, a separate group of 46 participants was recruited and presented with a paraphrase task that was analogous to the one in Experiment 1 (see bottom panel in Figure 1).[10] Specifically, they read each item in its entirety (the context as well as the sluiced question) and were then asked to paraphrase the ellipsis site using their own words. The within-item modal response for each of the 30 experimental items was selected and served as an answer choice in the forced-choice passage completion task in Experiment 2b.

6.2.3 Predictions

We expect the predictability of the sluiced questions in our materials to vary both within and across contexts. If felicitous sluicing depends on the salience of the relevant QUD in context, we would expect the acceptability scores from Experiment 2a to be positively correlated with the question-predictability scores collected in this experiment.

6.3 Results

Three participants were excluded for self-identifying as non-native speakers of English, and an additional 53 observations were excluded because they were submitted in less than 5000 ms (we assume that it takes at least 5 seconds to carefully read the prompt and all answer choices), leaving us with a total of 253 individual observations from 47 participants.

The results are shown in Figure 6. In each context, some questions were considered more likely than others. For example, given the context *Regarding Trump's impeachment,*

10 11 participants were excluded for either self-identifying as a non-native speaker of English or for failing to perform the task correctly. For example, several participants copied a random portion of the experimental prompt into the text box, which is behavior exhibited by automated MTurk bots.

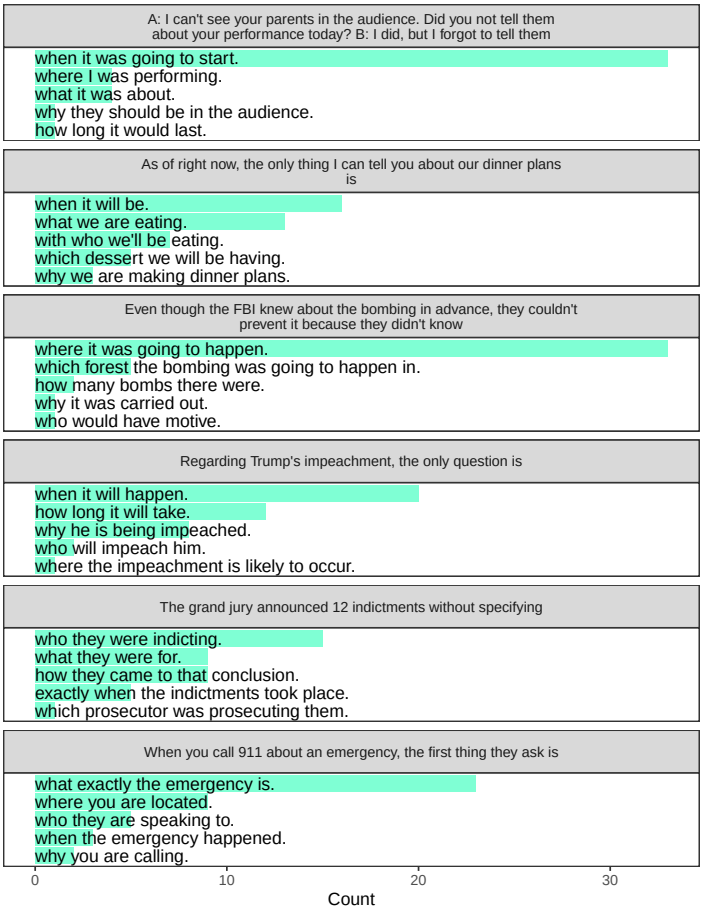


Fig. 6: Results from Experiment 2b. Bars indicate the number of times participants chose each question as the most likely continuation given the context.

the only question is..., the most likely continuation was the question *when it will happen*, where as the questions *who will impeach him* and *where the impeachment is likely to occur* were extremely unlikely. The key question for Experiment 2b was whether the variability in question predictability, summarized in Figure 6, is correlated with the acceptability of sluicing those questions, which was measured in Experiment 2a.

Figure 7 suggests that there is indeed a positive correlation: the more likely a question was to be selected in the forced-choice passage completion task, the higher its expected acceptability rating. To confirm whether this relationship is statistically significant, we conducted a hierarchical ordinal regression analysis with acceptability as a

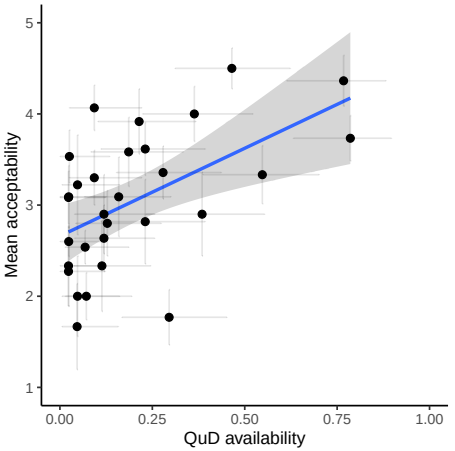


Fig. 7: Acceptability (Experiment 2a) as a function of QUD availability (Experiment 2b). Vertical error bars reflect Standard Errors; horizontal error bars show 95% exact confidence intervals.

categorical outcome variable and question predictability (i.e., the within-item proportion of trials on which the relevant question was selected in the passage completion task in Experiment 2b) as a population-level predictor. We additionally included by-item and by-participant group-level intercepts and slopes for question predictability. The results confirm that acceptability of sluicing does indeed increase significantly as a function of question predictability ($\Delta = 2, CI(\Delta) = [-0.23, 4.2], P(\Delta > 0) = 0.97$).

6.4 Discussion

The purpose of Experiment 2b was to test whether the acceptability of sluicing questions in nominal-antecedent contexts is correlated with the predictability of those questions given the context. To that end, we operationalized question predictability in terms of a forced-choice passage completion task in which participants were presented with the same contexts that featured the sluiced questions in Experiment 2a and selected the most likely continuation. Both the predictability of the questions and the acceptability of sluicing them exhibited variability and, as predicted, the two measures were positively correlated.

What we don't know at this point is whether question predictability is predictive of the acceptability of sluicing *per se* or whether it is more generally associated with the acceptability of the question regardless of whether it is sluiced or not. The reason for that is that Experiment 2a did not include unelided variants of the sluiced questions because

it is non-trivial to determine the meaning of those questions. Since we had to establish the meaning of each question for the passage-completion paradigm in Experiment 2b, however, we can now test the acceptability of the unelided variants to see whether the question predictability effect is specific to sluicing or not.

7 Experiment 3: acceptability including unelided variants

The goal of Experiment 3 was twofold: to replicate the findings from Experiments 2a and 2b; and to test whether the same pattern holds for unelided variants of the sluiced questions. We thus replicated the acceptability judgment task from Experiment 2a and included unelided variants of each item. Those unelided variants were determined by selecting the modal response from the norming experiment we conducted in preparation for Experiment 2b in which participants paraphrased the ellipsis site.

7.1 Methods

7.1.1 Materials

The materials were identical to those used in Experiment 2a except that unelided variants (see Figure 6) were added, as shown in (21).

(21) When you call 911 about an emergency, the first thing they ask is...

a. ...where (you are located).
b. ...why (you are calling).
c. ...what exactly (the emergency is).
d. ...who (they are speaking to).
e. ...when (the emergency happened).

The experimental items were interspersed with the same 12 filler items used in Experiment 2a (2:1 ratio), half of which were acceptable and half unacceptable.

7.1.2 Procedure & participants

We recruited 181 participants via `Amazon.com`'s Mechanical Turk. 19 of those participants were excluded for self-identifying as non-native speakers of English. The data

from the remaining 162 participants was further filtered to exclude 612 individual trials on which the response was submitted in less than 1000 ms,[11] leaving us with a total of 2916 individual observations.

Participants read each experimental item and judged it in terms of its acceptability, using a 5-point Likert scale ranging from "unacceptable" to "fully acceptable."

7.1.3 Predictions

We expect to replicate the results from Experiments 2a: the acceptability of sluiced questions should increase as a function of the degree to which those questions are predictable from the context. The key question is whether or not the unelided variants show the same effect. If the question predictability effect is specific to sluicing, we should expect to see a significant interaction whereby the effect is significantly less pronounced for unelided variants.

7.2 Results

The results, shown in Figure 8, suggest that the finding from Experiments 2a and 2b is replicated fully: the acceptability of sluiced questions (left panel) appears to increase with question predictability. The novel unelided variants, however, also exhibit a positive, albeit somewhat attenuated, relationship with question predictability. Since they are overall much more acceptable than their sluiced counterparts and are pushing up against the upper bound of the acceptability scale, the attenuation of the question predictability effect may be due to a ceiling effect.

To test these patterns statistically, we fit a hierarchical cumulative probit model with question predictability, ellipsis (treatment-coded), and their interaction as population-level effects.[12] We also added crossed group-level intercepts and slopes corresponding to all population-level effects for both items and participants (Barr et al., 2013). This analysis confirmed that elided variants showed a strong question-predictability effect ($\Delta = 2.01$, $CI(\Delta) = [0.93, 2.96]$, $P(\Delta > 0) = 1$) and that unelided variants were significantly more acceptable than their sluiced counterparts ($\Delta = 1.38$, $CI(\Delta) = [1.03, 1.74]$, $P(\Delta > 0) = 1$). The interaction between the two, however, was not significant: $\Delta = -0.73$, $CI(\Delta) = [-2.33, 0.91]$, $P(\Delta < 0) = 0.83$. In other words, the effect of question pre-

11 As before, we assume that it is not possible to carefully read and judge the experimental items in less than a second.

12 This type of model is particularly well-suited for accounting for the possibility of ceiling and floor effects (e.g., Fernandez, Liu, & Costilla, 2019).

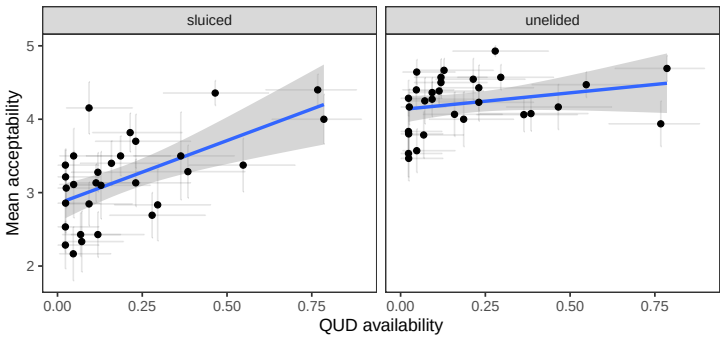


Fig. 8: Acceptability (Experiment 3) as a function of QUD availability (Experiment 2b). Vertical error bars reflect Standard Errors; horizontal error bars show 95% exact confidence intervals.

dictability was not limited to, or significantly more pronounced for, sluicing compared to unelided questions.

7.3 Discussion

The results from Experiment 3 confirm that the acceptability of our experimental items is affected by question predictability. They do not, however, rule out the possibility that the effect of question predictability affects unelided variants just as much as it does sentences that involve sluicing. While there is a numerical interaction between ellipsis and question predictability in the right direction (unelided variants were numerically affected less by question predictability), this effect did not reach statistical significance.

One potential concern is that the question predictability measure failed to identify many high-predictability items, leading to a lack of statistical power towards the upper end of the question-predictability scale (see Figure 8).[13] This is rooted in the way question predictability was operationalized in Experiment 2b: the forced-choice passage-completion paradigm yields a normalized probability distribution in which the scores of all questions in a given context sum to 1. As a result, high scores are only possible when all but one question are considered highly unlikely. It is therefore possible that a different operationalization of question predictability may lead to more statistical power and thus might have a better chance of detecting a sluicing-specific effect if one does, in fact, exist. As things stand, however, our results cannot rule out the possibility that the

13 Thanks to Dan Lassiter for pointing this out to us.

effect of question predictability reflects a general pragmatic effect that is not specific to sluicing.

8 General Discussion

8.1 Sluicing in the face of extreme lexical mismatches

In three experiments, we have explored the acceptability of sluicing under extreme mismatches between the elided material and its antecedent. A key finding is that such uses of sluicing can be highly acceptable despite the lexical mismatches they incur. In this section, we discuss the implications of this state of affairs for theories of sluicing. Consider the following examples, which were among the most highly rated items across the experiments reported:[14]

(22) a. Can I get a few autographs? —Sure, how many (do you want/need)?
b. I think the 49ers need a new quarterback. —Agreed, but who (should it be)?
c. Can I borrow your textbook over the weekend? —Why (do you want to borrow it)?

(23) a. I can't see your parents in the audience. Did you not tell them about your performance today? —I did, but I forgot to tell them where (it would take place).
b. When you call 911 about an emergency, the first thing they ask is where (you are located).
c. Regarding Trump's impeachment, the only question is when (he will be impeached).

The fact that sluicing is felicitous in these cases is problematic for all identity theories of which we are aware. For example, Merchant's (2001) purely semantic e-GIVENness constraint predicts them to be ungrammatical because the elided material and the antecedent do not entail each other. Rudin's (2019) purely lexico-syntactic formulation—which requires that identity only be maintained within the eventive core so as to capture the possibility of certain types of mismatches (e.g., modality, polarity, tense, and finiteness mismatches)—also makes the wrong predictions for these cases, since the operative lexical mismatches are located underneath the highest vP. Finally, both constraints imposed

14 As before, the material in brackets corresponds to the most common response as determined via the paraphrase task described in Section 6.2.2.

by Chung's (2006) hybrid identity proposal—e-GIVENness and No New Words—are violated in these examples. It is in fact hard to see how any approach could handle these cases and still be aptly characterized as an identity theory.

It is important to note that certain mismatches we have considered in this chapter also violate Chung's (2013) more limited syntactic identity account. According to this account, the No New Words constraint only applies to lexical elements that either assign case to the remnant wh-phrase or else determine the argument structure that the remnant participates in. These mismatches are problematic for Chung (2013) for two reasons. First, her account incorporates e-GIVENness, which is violated independently of the No New Words constraint and the "special heads" restriction. Second, while some of the mismatches that proved to be highly acceptable are permitted under Chung's limited syntactic identity condition, some are clearly not exempt from it. For example, in (22b) the sluicing remnant *who* serves as an argument to the elided predicate and thus requires an identical correlate in the antecedent clause, which it does not have.[15]

As a result, the example should be ungrammatical according to Chung (2013), but it was judged to be highly acceptable nonetheless.[16] Whereas there has been a tradition of theorizing about ellipsis that revolves around the incremental fine-tuning of the definition of identity in order to capture increasingly large sets of otherwise problematic observations, the examples in (22) and (23) do not lend themselves to this strategy, as they represent instances of a heterogeneous set of mismatches that cannot be captured in a small number of exceptions. This problem is exacerbated by the fact that our experiments did not just reveal highly acceptable mismatches, but also completely unacceptable ones as well as marginally (un)acceptable cases. Any amendments to the definition of identity that would capture the acceptable cases would thus likely

15 An anonymous reviewer suggests that the sluiced question could be analyzed as *who do they need*, which could be seen to preserve identity with the antecedent. Evidence against this possibly is provided by the fact that participants in the cloze-completion experiment overwhelmingly paraphrased the question as *who should it be*. This interpretation corresponds to our own intuitions for this example.

16 An anonymous reviewer points out that nominals follow the same pattern as verbs and adjectives in requiring that prepositions that mark arguments occur overtly in either the antecedent or the ellipsis site. That is, example (i) follows the pattern we saw in (3)–(4) for verbs and adjectives.

(i) a. A: You aren't going anywhere. We're expecting a delivery.
b. B: Of what?
c. B: # What?

The challenge for any theory, therefore, is explaining why nominals can give rise to a wide array of inferential referents that can support subsequent sluicing, but at the same time require that case-marking prepositions be overt in either the antecedent or ellipsis clauses (see Jäger, 2001, and Barker, 2013, for attempts to formalize that requirement within a referential framework, following an idea that was to our knowledge first articulated in Ginzburg, 1992, p. 146).

overgenerate with respect to the marginal and fully unacceptable cases. The next section considers the implications of the gradience in acceptability that our experiments revealed in more detail.

While we have focused our attention thus far on the implications of our results for theories of ellipsis that posit some type of syntactic identity constraint, we also note that there are theories that do not posit the existence of structure at the ellipsis site and, in being governed instead by identity conditions that are semantic, are compatible in principle with some aspects of our findings (e.g., Ginzburg & Sag, 2000; Culicover & Jackendoff, 2005, 2012). However, no such approach, to our knowledge, has engaged with the full set of effects identified here, including the extent to which the recovered meaning of the elided material can differ from the meaning of the antecedent, and the high degree of variability we find across items. In our view, any successful theory will need to engage with the referential system of languages to capture the many similarities between ellipsis and other discourse-referential devices (see Kehler, 2019b; Poppels, 2020, for a list of shared diagnostic properties), to account for the various factors that affect the salience of discourse representations and ultimately the success of subsequent reference, and to leverage the independently-motivated inferential machinery operative across the referential domain.

8.2 Variable acceptability

Besides the existence of highly acceptable mismatches as exemplified in (22) and (23), the considerable degree of variance in acceptability across items constitutes another key finding of our investigation. This finding is inherently problematic for identity theories, which aim to classify sentences as either grammatical or ungrammatical. Referential theories of ellipsis, on the other hand, naturally predict gradience since the theoretical constructs they leverage, such as salience and plausibility, are themselves inherently gradient in nature. According to such theories, any deviation from the meaning introduced by the antecedent requires some degree of inference, and the ability to draw the relevant inference—and thus whether or not the use of ellipsis is felicitous in such contexts—depends on factors that are themselves poorly understood to date. The challenge for advocates of referential theories is therefore to identify testable predictions about the acceptability of inferentially resolved ellipsis.

Any such theory would have to contend with the strongly contrasting judgments witnessed in minimal pairs such as (24) in Experiment 2a.

(24) a. Regarding Trump's impeachment, the only question is when (he will be impeached).

b. Regarding Trump's impeachment, the only question is who #(will vote for it).

In Experiment 2b we attempted to explain the gradience in acceptability we had identified in Experiment 2a based on the hypothesis that more predictable questions would be easier to infer. We operationalized question predictability in terms of a forced-choice passage completion task. This task succeeded in capturing a decent amount of variability, but it was inherently limited in that it failed to identify but a few highly predictable questions in the set of experimental items. As a result, there was limited statistical power towards the upper end of the question-predictability scale since most measures were clustered around the lower end of the scale. Despite this shortcoming, question predictability was significantly correlated with the acceptability of the sluiced variants. This predictability effect did not, however, affect sluiced variants significantly more than their unelided variants, and thus we do not have evidence that the effect is ellipsis-specific.

While it is possible that an improved operationalization of the concept of question predictability could reveal a sluicing-specific component (after all, the relevant interaction term was trending in the right direction numerically), it is also possible that question predictability simply does not play a role in the gradience in acceptability associated with nominal-antecedent sluicing. In fact, there are ellipsis-independent reasons to believe that the predictability or salience of the intended discourse referent is not sufficient for ensuring that the use of referring expressions is felicitous. Consider the following example, due to Barbara Partee:

(25) a. I dropped ten marbles and I found all but one of them. It must be under the sofa.
b. I dropped ten marbles and I found only nine of them. # It must be under the sofa.

In (25a), *it* felicitously refers to the tenth marble, which is explicitly introduced into the discourse by the antecedent NP *one of them*. By contrast, the context in (25b) does not explicitly mention the missing marble and it must therefore be inferred. The fact that this inferential use of the pronoun *it* is infelicitous, however, is surprising given the fact that several aspects of the context conspire to make it maximally salient and predictable: not only does the context plausibly raise the question *Did you find the marbles you dropped?*, the focus construction *only nine of them* specifically shifts attention to the last missing marble. As a result, if inferential reference resolution were reducible to the accessibility of the intended referent, we should expect the use of *it* to be perfectly felicitous in this context, but it clearly is not. We suspect that the pronoun *it* requires not just a salient referent, but further presupposes that that referent has previously been established as topical, and that explicitly denoting it in a preceding NP goes a long way toward serving that function.

If sluicing is a form of discourse reference, it is possible that inferential sluicing of the kind we examined in this chapter is likewise not reducible to the accessibility of the intended referent. Fully understanding the gradience associated with inferentially resolved sluiced questions will require a more detailed understanding of the intersection between concepts such as salience, topichood, and predictability.

9 Conclusion

Across a series of three experiments, we investigated the possibility of inferentially resolved sluicing. We demonstrated that sluicing can be highly acceptable even when the elided material deviates substantially from its antecedent, both semantically and syntactically. We further found considerable variability across items, which raises an important new challenge for all existing theories. While the existence of such gradience is inherently consistent with referential theories of sluicing, our investigation highlights the challenge of explaining this variability on an item-by-item basis. identity theories, on the other hand, are challenged in a more fundamental way by our findings. These findings suggest that the No New Words constraint does not generalize beyond the narrow set of preposition-stranding cases that originally motivated the proposal, and further suggest that semantic identity conditions like e-GIVENness, which continue to play a central role in most identity proposals to date, are overly restrictive. Most importantly, the heterogeneous class of mismatches our investigation examined does not lend itself to the fine-tuning approach that is prevalent among identity theorists.

Acknowledgment: We are grateful to two anonymous reviewers for helpful comments on an earlier draft, and also thank Andreas Konietzko and the participants of the Information Structure and Ambiguity workshop at the University of Tübingen, members of the UCSD Semantics Babble reading group, and audience members at a talk given at the 94th Annual Meeting of the Linguistics Society of America for useful feedback. We also gratefully acknowledge support from NSF grant BCS-1456081 to Roger Levy.

A Experimental items used in Experiment 1

A.1 "Inference" items:

(26) a. Fan: Can I get a few autographs?
Manager: Sure, how many?

b. Child: Can we get something to eat that I like?
Parent: It depends on what.
c. Fan: I really want to hear at least one Justin Bieber song before the night is over.
DJ: OK, I just have to figure out when.
d. Guest: Can I get another drink on the house?
Waiter: Tell me which one and I'll find out.
e. A: Can I borrow your textbook over the weekend?
B: Why?
f. A: Do you mind if I put on some music?
B: What kind of music?
g. Friend 1: Can we see some sort of performance from you at the party tonight?
Friend 2: What type of performance?
h. Parent: Could you unload the dishwasher?
Child: When?
i. Professor: Could you sign up for my class next quarter?
Student: Which one?
j. A: Would you mind turning the TV off?
B: Why?
k. Spectator: Can I please see a card trick?
Magician: I'm sorry, I don't know how.
l. Husband: I'd love to watch the ballgame on my phone while we're out.
Wife: No problem, I'll just have to remember how.
m. Customer: Are those the steaks that are on sale?
Butcher: Yes, sir. How many?
n. A: Apparently, the FBI knew about the bomb threat in advance.
B: But did they also know when?
o. Player 1: Did you not tell your friends about the game today?
Player 2: I did, but I forgot to tell them where.
p. Police officer: And shortly after 3 pm is when you heard the explosion?
Witness: Yes, but I couldn't tell where.
q. Football fan 1: I think the 49ers need a new quarterback.
Football fan 2 Agreed, but who?
r. Sales assistant: What's a phone number at which we can reach you?
Customer: Why?
s. Analyst: If this company is to thrive in the long run, it needs a more aggressive CEO.
HR manager: Well, tell me who and I'll take care of it.

t. Mother: I think our son would do much better with a different math teacher.
Father: That depends on who.

A.2 "Classic" items

Items consisted of pairs of dialogues that either did or didn't include the material shown in square brackets.

(27)
a. Claire: Joe is jealous [of someone].
Jessica: Do we know who?
b. Widower: Susan was murdered [by someone].
Son in law: Have they revealed who?
c. Nanny: Last night he was very afraid [of something].
Mother: Did he say what?
d. A: Mary was flirting [with more than one guy].
B: Did you see who?
e. Mr. Jones: We're donating our car [to some organization].
Mr. Smith: Have you decided which organization?
f. Father: Sarah is reading a novel [by a Norwegian author].
Teacher: Do you know who?
g. Brother: Dad said he's angry [with someone].
Sister: Did he mention who?
h. Press secretary: The board of directors had a meeting [with a client].
Reporter: Let me guess: we're not allowed to know which client.
i. A: The reporter asked at least 20 questions [about one topic].
B: Which topic?
j. A: The boy is visibly happy [about something].
B: True, but it isn't clear what?

B Experimental items used in Experiments 2 and 3

Experiment 3 additionally included the unelided variants identified in the norming experiment described in Section 6.2.2.

(28) A: I can't see your parents in the audience. Did you not tell them about your performance today?
B: I did, but I forgot to tell them...
a. ...where.

b. ...when.
c. ...why.
d. ...how long.
e. ...what about.

(29) As of right now, the only thing I can tell you about our dinner plans is...

a. ...when.
b. ...what.
c. ...with who.
d. ...which dessert.
e. ...why.

(30) When you call 911 about an emergency, the first thing they ask is...

a. ...where.
b. ...why.
c. ...what exactly.
d. ...who.
e. ...when.

(31) The grand jury announced 12 indictments without specifying...

a. ...who.
b. ...what for.
c. ...exactly when.
d. ...which prosecutor.
e. ...how.

(32) Regarding Trump's impeachment, the only question is...

a. ...when.
b. ...who.
c. ...why.
d. ...where.
e. ...how long.

(33) Even though the FBI knew about the bombing in advance, they couldn't prevent it because they didn't know...

a. ...where.
b. ...who.
c. ...how many.
d. ...why.
e. ...which forest.

References

AnderBois, S. (2010). Sluicing as anaphora to issues. In *Proceedings of Semantics and Linguistic Theory* (Vol. 20, pp. 451–470).

AnderBois, S. (2014). The semantics of sluicing: Beyond truth conditions. *Language*, *90*(4), 887–926.

Barker, C. (2013). Scopability and sluicing. *Linguistics and Philosophy*, *36*(3), 187–223.

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.

Barros, M. (2014). *Sluicing and identity in ellipsis* (Doctoral dissertation, Rutgers University).

Beecher, H. (2007). Pragmatic Inference in the Interpretation of Sluiced Prepositional Phrases. In *Proceedings of the Fifth University of Cambridge Postgraduate Conference in Language Research* (pp. 9–16). Citeseer.

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

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the R package brms. *The R Journal*, 1–15.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Chung, S. (2006). Sluicing and the lexicon: The point of no return. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 31, pp. 73–91). Berkeley, CA.

Chung, S. (2013). Syntactic identity in sluicing: How much and why. *Linguistic Inquiry*, *44*(1), 1–44.

Chung, S., Ladusaw, W. A., & McCloskey, J. (1995). Sluicing and logical form. *Natural Language Semantics*, *3*(3), 239–282.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler Syntax*. Oxford University Press on Demand.

Culicover, P. W., & Jackendoff, R. (2012). Same-except: A domain-general cognitive relation and how language expresses it. *Language*, *88*(2), 305–340.

Dalrymple, M., Shieber, S. M., & Pereira, F. C. (1991). Ellipsis and higher-order unification. *Linguistics and Philosophy*, *14*(4), 399–452.

Fernandez, D., Liu, I., & Costilla, R. (2019). A method for ordinal outcomes: The ordered stereotype model. *International Journal of Methods in Psychiatric Research*, *28*(4).

Fiengo, R., & May, R. (1994). *Indices and identity*. MIT press.

Fu, J., Roeper, T., & Borer, H. (2001). The VP within Process Nominals: Evidence from Adverbs and the VP Anaphor Do-So. *Natural Language & Linguistic Theory*, *19*(3), 549–582.

Ginzburg, J. (1992). *Questions, queries and facts: A semantics and pragmatics for interrogatives* (Doctoral dissertation, Stanford University).

Ginzburg, J. (2012). *The interactive stance*. Oxford University Press.

Ginzburg, J., & Miller, P. (2019). Ellipsis in head-driven phrase structure grammar. In J. van Craenenbroeck & T. Temmerman (Eds.), *Handbook of ellipsis*. Oxford: Oxford University Press.

Ginzburg, J., & Sag, I. A. (2000). *Interrogative investigations*. Stanford: CSLI publications.

Hale, K., & Keyser, S. J. (1993). On Argument Structure and the Lexical Expression of Syntactic Relations. In K. Hale (Ed.), *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger* (pp. 53–109). MIT Press.

Hankamer, J., & Sag, I. A. (1976). Deep and surface anaphora. *Linguistic inquiry*, *7*(3), 391–428.

Hardt, D. (1993). *Verb phrase ellipsis: Form, meaning, and processing* (Doctoral dissertation, University of Pennsylvania).

Hardt, D., Anand, P., & McCloskey, J. (2020). *Sluicing: Antecedence and the landscape of mismatch*. Talk presented at Experimental and Corpus-based Approaches to Ellipsis.

Jäger, G. (2001). Indefinites and sluicing. A type logical approach. In *Proceedings of the 13th Amsterdam Colloquium*.

Johnson, K. (2001). What vp ellipsis can do, and what it can't, but not why. In *The Handbook of Contemporary Syntactic Theory* (pp. 439–479). Wiley Online Library.

Kehler, A. (1993). The effect of establishing coherence in ellipsis and anaphora resolution. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (pp. 62–69).

Kehler, A. (2016). On QUD-based licensing of strict and sloppy ambiguities. In *Semantics and linguistic theory* (Vol. 25, pp. 512–532).

Kehler, A. (2018). *Prominence in a referential theory of vp-ellipsis*. Keynote Talk Presented at Prominence in Language 2018, Cologne, Germany.

Kehler, A. (2019a). Ellipsis and discourse. In J. van Craenenbroeck & T. Temmerman (Eds.), *Handbook of ellipsis*. Oxford University Press.

Kehler, A. (2019b). Ellipsis and discourse. In J. van Craenenbroeck & T. Temmerman (Eds.), *The Oxford Handbook of Ellipsis* (pp. 314–341). Oxford: Oxford University Press.

Kehler, A., & Ward, G. (2007). Event reference and morphological transparency. In *Proceedings of the 35th western conference on linguistics (wecol-07)* (pp. 115–127). University of California, San Diego.

Kertz, L. (2013). Verb phrase ellipsis: The view from information structure. *Language*, *89*(3), 390–428.

Lipták, A. (2015). Identity in ellipsis: An introduction. *Lingua*, (166), 155–171. doi:10.1016/j.lingua.2015.08.003

Merchant, J. (1999). E-type A'-traces under Sluicing. In *The Proceedings of the Seventeenth West Coast Conference on Formal Linguistics* (Vol. 30, pp. 478–492).

Merchant, J. (2001). *The Syntax of Silence: Sluicing, islands, and the Theory of Ellipsis*. Oxford University Press on Demand.

Merchant, J. (2008). An Asymmetry in Voice Mismatches in VP-Ellipsis and Pseudogapping. *Linguistic Inquiry*, *39*(1), 169–179.

Merchant, J. (2013a). Polarity items under ellipsis. In L. L.-S. Cheng & N. Corver (Eds.), *Diagnosing syntax*. Oxford University Press Oxford.

Merchant, J. (2013b). Voice and ellipsis. *Linguistic Inquiry*, *44*(1), 77–108.

Miller, P., & Hemforth, B. (2014). *Verb phrase ellipsis with nominal antecedents*.

Miller, P., & Pullum, G. K. (2013). Exophoric VP Ellipsis. In P. Hofmeister & E. Norcliffe (Eds.), *The Core and the Periphery: Data-driven Perspectives on Syntax Inspired by Ivan A. Sag* (pp. 5–32).

Poppels, T. (2020). *Towards a referential theory of ellipsis* (Doctoral dissertation, University of California San Diego).

Potsdam, E. (2007). Malagasy sluicing and its consequences for the identity requirement on ellipsis. *Natural Language & Linguistic Theory*, *25*(3), 577–613.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*(6), 1–69. Published version of draft circulated in 1996 and amended in 1998. doi:10.3765/sp.5.6

Ross, J. (1969). Guess who? In *Papers from the fifth regional meeting of the Chicago Linguistic Society.* (pp. 252–286).

Rudin, D. (2019). Head-based syntactic identity in sluicing. *Linguistic Inquiry*, *50*(2), 253–283.

Sag, I. A. (1976). *Deletion and logical form* (Doctoral dissertation, Massachusetts Institute of Technology).

Tanaka, H. (2011a). Syntactic identity and ellipsis. *The Linguistic Review*, *28*(1), 79–110.

Tanaka, H. (2011b). Voice mismatch and syntactic identity. *Linguistic Inquiry*, *42*(3), 470–490.

Ward, G., Sproat, R., & McKoon, G. (1991). A pragmatic analysis of so-called anaphoric islands. *Language*, *67*(3), 439–474.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison Wesley.