The Santa Cruz Sluicing Dataset

Pranav Anand
University of California Santa Cruz

Daniel Hardt
Copenhagen Business School

James McCloskey
University of California Santa Cruz

This paper describes a new research resource – a searchable database of 4700 naturally occurring instances of sluicing in English, annotated so as to shed light on the questions which have shaped research on ellipsis since the 1960's. The paper describes the dataset and how it can be obtained, how it was constructed, how it is organized, and how it can be queried. It also highlights some initial empirical findings, first describing general characteristics of the data, then focusing more closely on issues concerning antecedents and possible mismatches between antecedents and ellipsis sites.[1]

1. Introduction. Ellipsis is a pervasive and mysterious aspect of human language, one whose effects are felt in every subdomain of grammar and one which every branch of the language sciences must come to terms with. This paper introduces a new research resource devoted to this important phenomenon – a dataset of several thousand naturally occurring instances of sluicing (prepared by the Santa Cruz Ellipsis Project), annotated so as to shed light on the issues that have driven research on ellipsis since the 1960's. In its size and in the sophistication of its annotation scheme, the corpus is, we believe, unprecedented and our aim in developing it has been to make available to the various research communities who must care about ellipsis a robust evidential basis for theory testing and, perhaps more importantly, an impetus for the asking of new kinds of questions. Here, we introduce the dataset and its principal properties, describing its construction and illustrating how it might be useful by presenting some initial findings that emerge from it. We leave the work of theoretical interpretation for another occasion.

Our ultimate concern is with ellipsis in general, but our initial focus was on English and on sluicing – as in (1):

(1) She will resign, but we don't know when.

In sluicing, all but the interrogative phrase of a content question is elided. We chose sluicing as initial target because it is widely attested across languages (making it a good starting point from which to extend beyond English), because in English it is widely used in many registers and

[1]The research reported here was supported by funding from the Academic Senate of UC Santa Cruz, from The Humanities Institute of UC Santa Cruz, and from the National Science Foundation via Award Number 1451819: *The Implicit Content of Sluicing* (pi Pranav Anand, co-pis James McCloskey, and Daniel Hardt). The project would have been impossible without the perceptiveness and commitment of our undergraduate annotators: Brooks Blair, Jacob Chemnick, Charlotte Daciolas, Jasmine Embry, Jack Haskins, Anny Huang, Zach Lebowski, Lily Ng, Lyndsey Olsen, Reuben Raff, and Serene Tseng. Particularly important contributions were made by our lead annotators – Rachelle Boyson, Mansi Desai, Chelsea Miller, Lydia Werthen, and Anissa Zaitsu. Our graduate student research assistants also made crucial contributions: Kelsey Kraus, Margaret Kroll, Deniz Rudin, and Bern Samko. Beyond the project itself, many colleagues have provided advice and support which we appreciate – Sandy Chung, Vera Gribanova, Kyle Johnson, Jason Merchant, and Tim Stowell in particular.

genres (so a relatively large corpus could be assembled), because it is well studied (we therefore had the ingredients for a useful annotation system), and because it interacts in interesting ways with many other important aspects of form and interpretation – questionhood, the dynamics of discourse, the organization of lexical information, the representation of implicit content, the difference between root and embedded structures, the syntax of WH-movement, and much else.

The dataset consists of 4700 instances of sluicing in English, each taking the form of a short text annotated for syntactic, semantic, and pragmatic characteristics.[2]

Each example includes a substantial context window (preceding and following), by way of which properties of its discourse context can be scrutinized. This information is crucial, since ellipsis is richly and subtly sensitive to the context of use – in acceptability and in interpretation. It is not difficult for a trained investigator in syntax or semantics to invent revealing examples in isolation; what often matters most for ellipsis, though, is not the example in isolation but the contexts, sometimes large and often intricate, in which it might be used. Conjuring up such contexts is not easy, but for the central questions concerning ellipsis, they are crucial. Our project therefore grows out of the conviction that corpus work can be of central importance in deepening our understanding of ellipsis and further that the time is now right to turn to this methodology on a larger scale and in a more systematic way than has been customary. With currently available tools very large datasets can be constructed which include discourse contexts and which can be mined to uncover new patterns and to test hypotheses – all on a scale far beyond what is possible with individually constructed examples.

The dataset we describe can be browsed here and can be downloaded here. The download contains, for each example, a pair of plain text files – the first containing the example itself with its discourse context, the second containing the annotations. Since each example has a unique numerical identifier, that number names both files. Example 100640, then, is characterized by the combination of two files – 100640.txt (the example itself with its context window) and 100640.ann (its annotation). The view presented on the annotation interface here combines the information in this pair of files into a single visual representation.

Since all of the data (text and annotations) is represented in plain text format, any string based search tool can be used to query it. In addition, the 1.3 data release includes a simple Python script to walk through and examine the data. The key program is *explorer.py*, which walks through a plain text file consisting of a list of jsons, each of which contains the data for one example. The script selects elements that match a user's search query and prints them out as a static *.html* file. The release contains some sample query files and also instructions about how to alter, extend, and customize queries.

2. DATA. Our data comes almost exclusively from the New York Times subcorpus of the English Gigaword (Second Edition) corpus (Graff et al. 2005). We first parsed the corpus with the Stanford parser (Klein & Manning 2003) and then used TGrep2 (Rohde 2005) to extract all verb phrases whose final child was a WH-phrase. That yielded 5100 verb phrases, which were then manually culled to eliminate false positives. That process in turn yielded 3374 true instances of sluicing in nonroot settings. As a check, all 52,000 WH-phrases in a random 80th of the NYT subcorpus were manually examined. This procedure yielded just one additional sluice and provided some grounds

[2]We also provide 1200 unannotated examples. These are of two types – examples which sufficiently resemble sluicing that they turned up as false positives in our searches, and instances of sluicing which, for mostly technical reasons, proved too difficult to annotate within our system.

for confidence that our procedures successfully identified virtually all instances of embedded sluicing in the subcorpus.

Root sluices were harder to identify, since the structures provided by the parser for them were too inconsistent to support automated searching. To find such cases, we first isolated all root WH-questions – some 91,000 examples (including many false positives). These examples were examined manually and from among them, 1289 examples of root sluices were identified for annotation. To fill out this sample, we added 37 examples from other written sources, as we happened on them. These are our 4700 annotated examples – 62.4% embedded and 27.6% root sluices.

3. ANNOTATION. The annotation system that we brought to bear on this data was shaped by a set of partially conflicting goals. It had to strike a balance, in the first place, between theoretical sophistication and usability – for end users and for annotators alike. But a scheme which sought to avoid all theoretical commitment would be of little use. We therefore drew heavily on the existing theoretical literature on sluicing in designing our protocols. At the same time, however, the scheme had to be sufficiently catholic to be useful to researchers with a variety of purposes in mind and of different theoretical persuasions.

For these reasons and others, we elected not to do our annotations on syntactic or semantic representations; any representational system we chose would necessarily privilege a particular theoretical point of view. Most of our features, therefore, simply refer to spans of text. Some repercussions of that choice will be considered below.

The best way to understand our annotation system is to consult our coding manual, which is available for download. Here we provide an overview of its central features.

Each example is annotated with five obligatory tags: (i) the ANTECEDENT, (ii) the WH-remnant, (iii) a plain text paraphrase of the elided content, (iv) the MAIN PREDICATE of the antecedent clause, and (v) the CORRELATE of the WH-remnant, if there is one. The correlate and the WH-remnant are both tagged with several taxonomic features, including syntactic and semantic type. Consider (2), for example:

(2) Brady said the new approach saves time, but she didn't know how much. [100452]

Here, the two word span *how much* would be identified as the WH-remnant and the five word span *the new approach saves time* as the antecedent. The correlate is the single word span *time* and the plain text paraphrase of the implicit content will be *the new approach saves*. The main predicate of that clause is the verb *saves*. The semantic type of the WH-remnant is DEGREE and the semantic type of the correlate is MASS/RANGE.

We found that a context window radius of five sentences was usually sufficient, and sometimes necessary, in determining the intended antecedent and how it related to the ellipsis site. The task of identifying an appropriate antecedent was not always straightforward and often required an understanding of the larger structure of the text, particularly with respect to questions under discussion in the sense of Roberts 2012. Very occasionally it was necessary to resort to the full newspaper article (most can still be found online with some patience) in order to be confident about the intended interpretation.

Examples which seem to lack an antecedent can be found by searching for the tag 'missing antecedent' (MISSINGANTE). Root sluices are identified by a tag of that name and embedded sluices can be identified by searching for the tag QEMBEDDER. In the case of (2), for instance, the

QEMBEDDER attribute has the value *know*. This tag therefore provides information simultaneously about which sluices are embedded and about the range of interrogative embedding predicates in our materials.

Once the antecedent is identified, it can be copied and modified as necessary into the ellipsis site and the important task of identifying mismatches (in form and in interpretation) between the antecedent and the elided content can be tackled. Mismatches are categorized by way of a set of binary features indicating morphological mismatches (e.g. Case), syntactic mismatches (finiteness, polarity, syntactic category of the antecedent and so on) along with semantic mismatches (tense, indexicality, modality, polarity). Two tags on the antecedent serve to mark interpretive differences between antecedent and elided content. The E-TYPE tag marks indefinite material in the antecedent which is interpreted anaphorically in the ellipsis site. In (3), for example, the nominal *one of the kids* will be so tagged, since it can be interpreted as the definite *that kid* in the ellipsis site.

(3) For some reason or other that is one of the kids jumping out at me. And I don't know why. [15397]

IGNORE marks material which is semantically active in the antecedent but has no counterpart in the ellipsis site – parenthetical material, additive particles, focus particles such as *only*, and *even*, an interestingly large range of adverbial expressions, among others.

The crucial tag NEW WORDS, by contrast, identifies cases in which the interpretation of the ellipsis implies the presence of a lexical item which has no counterpart in the antecedent. Whether or not this possibility exists has been a central concern in discussions of sluicing.

Certain other tags apply to global properties of examples. We do not, for instance, assume that every expression which appears in a corpus is *ipso facto* well formed. Each example is therefore rated on a three point scale of acceptability in context (Low, Medium, or High). In the end, 178 of our 4700 examples (3.8%) were judged to be either moderately or severely unacceptable by annotators. This information is obviously crucial for any theoretical conclusions that one might want to draw from our materials.

A number of other global tags are worth mentioning here:

ISLAND: identifies examples which could be relevant for the debate about whether or not sluicing amnesties island effects (Ross 1969, Chung et al. 1995, Romero 1998, Merchant 2001, Barros et al. 2014 among many others).

PROBLEMATIC: tags examples which are difficult to appropriately annotate within the terms of our scheme. 31 examples are so tagged; they constitute a treasure trove of challenging analytical puzzles.

COOL: marks examples which annotators found interesting, or unusual, for one reason or another. This group of 213 examples is also a rich source of intriguing puzzles and interactions.

Given space limitations, this overview cannot be much more than a taster for the full range of phenomena exposed and searchable in our dataset. Additional tags will be discussed in what follows, but those who wish to exploit the full potential of our dataset should consult the full annotation manual.

3.1. THE PEOPLE AND THE PROCESS. Our frontline annotators were undergraduate students in the linguistics program at UC Santa Cruz. Annotators were recruited to the project on the basis

of interest and of having completed, and done well in, at least intermediate courses in syntax and semantics. Their preparation in course work meant that they could handle the technicalities of annotation. But they came to the work without precommitment to any theoretical point of view and brought to it a useful iconoclastic glee in the finding of difficult cases and problematic counterexamples. Their work was overseen at every point by graduate student research assistants and by faculty PIs.

Annotation was conducted on the BRAT web based annotation tool (see Stenetorp et al. 2012), modified in various ways (in particular to accept and display a free text paraphrase of elided content). Unlike some other annotation tools, BRAT does not alter the form of the text being annotated (it is a 'stand off' annotation tool in which the annotation content is stored separately from the target text). This choice made the calculation of inter annotator agreement rates more straightforward; it also means that those who, for whatever reason, do not want to work with the annotations we offer can still easily investigate the content of the source texts alone.

Development of the annotated dataset proceeded in two phases. In the first, over three years, our undergraduate assistants annotated all of the examples identified as potential instances of sluicing. Each individual annotated on the order of 40 examples per week, and each example was annotated independently by at least two individuals. In weekly meetings moderated by graduate student lead annotators and a PI, difficult cases were adjudicated and feedback was provided by the annotators about the effectiveness and usability of the annotation protocols. The coding manual was modified as work proceeded and as difficulties were encountered and resolved. In cases of strong disagreement among annotators, all competing analyses were maintained. In general though, discussion at the weekly meetings tended to converge on agreement around a single 'best annotation' for each example.[3]

In the second phase, three lead annotators (master's students who had themselves been front line annotators), revisited the entire corpus in a second series of weekly meetings under the supervision of one of the PIs. The entire corpus of annotations was revised, to comply with the policies and guidelines of the final annotation scheme and to adjudicate remaining disagreements. In this process, the lead annotators had access to, and made use of, all of the alternative annotations that had emerged in phase one.

The data we report here reflects this final round of reconsideration and discussion, but all rounds are preserved for analysis. Each of the annotations offered in the dataset, then, even the most routine, has been scrutinized by at least five project members. Nonroutine cases have been examined and discussed by between 7 and 10 members of the project – undergraduate students, graduate students and faculty PIs.

3.2. GUIDELINES, POLICIES, AND HARD CHOICES. One of the principal goals of our project was to document as fully as possible the range of allowed mismatches in meaning and in form between an antecedent clause and the elided clause in sluicing. Identifying such mismatches was the third step in the annotation process for each example, after the WH-remnant and the antecedent had been identified. This done, antecedent and paraphrase could be compared and mismatches documented.

Clearly it is crucial for this assessment that the paraphrases provided for elided content be consistent across annotators and annotations. Halfway through the development cycle, technical

[3]There were, of course, real ambiguities as well. We deal with these by duplicating the relevant example and providing two distinct annotations.

modifications to the annotation software gave annotators the ability to copy the antecedent into the ellipsis site and then alter it to the extent needed to accurately represent the interpretation of the sluice. This technical innovation simplified the annotators' task (by freeing them from the obligation of constructing a paraphrase de novo) and led to greatly improved levels of inter annotator agreement.[4]

A set of guidelines was put in place to maximize consistency in this process and eliminate irrelevant differences (laying out, for example, exactly how definite interpretations of E-TYPE pronouns should be rendered in the paraphrase). The vast majority of such policies regulate what are fundamentally stylistic matters and involve arbitrary choices. There was one choice that had to be made, however, whose implications go beyond the stylistic and which requires more discussion. A striking property of sluicing is that for a given case it is often possible to identify either a relatively small antecedent or a more inclusive one. The differences in meaning entailed by different choices can be very slight; but in syntactic terms the choices are often starkly different. Consider (4) – part of a discussion of the early days of radio.[5]

(4) There was always something new … improved equipment, innovative means of transmission, original shows coming down the network line from New York and Chicago and above all, the knowledge that [thousands upon thousands of people] clustered around a box that sat like a shrine in their living rooms, [listening]. It didn't really matter to what [those thousands and thousands of people WERE listening]. [36225]

The paraphrase offered in (4) (the 'official' agreed upon annotation) presupposes a relatively small and apparently discontinuous antecedent (*thousands and thousands of people listening*). But an alternative annotation would identify a larger antecedent, as in (5):

(5) It didn't really matter to what [those thousands upon thousands of people clustered around that box, listening].

The sense communicated by (5) is barely distinguishable, in context, from that communicated by (4). This is because in (4) the content of the more inclusive clause (*thousands of people clustered around the box …*) is smuggled into the implicit restrictor of the demonstrative determiner *those* so that the entire phrase means something like *those thousands and thousands of people who clustered around a box that sat like a shrine in their living rooms*. The larger and more verbose paraphrase in (5) involves less tampering with the antecedent but it also presupposes a severe (but amnestied) violation of the adjunct island condition. For this case, annotators were in no doubt that (4) was the more appropriate annotation. But that choice brings its own implications, since the paraphrase, to ensure wellformedness, must include a lexical item (the copula) which has no counterpart in the antecedent.

The policy we adopted for such cases is that annotators should select the smallest antecedent consistent with an accurate rendering of the meaning of the elided clause – a convention we call ANTECEDENT MINIMALITY. This guideline, favoring (4) over (5), proved extremely helpful in

[4]For detailed discussion of inter annotator agreement rates and other issues considered only briefly here, see Anand & McCloskey 2015.

[5]In (4) and in other examples cited, we surround the antecedent with square brackets, and the paraphrase of elided content appears also within square brackets and in a gray font. The unique numerical identifier for the example is also given. Each such identifier is a live link to the view of the example on the annotation interface.

ensuring regularity and consistency. But it is not, of course, a theoretically innocent choice. And while for the particular case of (4) the choice seems fairly clear, other cases are more difficult to adjudicate. Consider (6) for instance.

(6) "We were very concerned that [the mortality pattern] seemed [to be so abrupt and sudden from women], but without research," she said, "we did not know why [the mortality pattern WAS so abrupt and sudden from women]." [57485]

Here too there is a choice to be made between a larger antecedent – one including the verb *seem* – and the smaller antecedent annotators actually identified, consistent with the guideline of ANTECEDENT MINIMALITY. This is the paraphrase given in (6), which assumes a smaller (and again discontinuous) antecedent (*the mortality pattern to be so abrupt and sudden from women*) and a consequent mismatch in form between antecedent and paraphrase: infinitival *to* in the antecedent corresponding to tensed *was* in the paraphrase. The judgment call required here is delicate: does the meaning of the elided question include the subtle evidential component contributed by *seem*? The answer implied by the final annotation (arrived at after considerable discussion) is that it does not. This conclusion is neither unreasonable nor obviously correct.

We air these issues for two reasons. First, it is important that users of our dataset be aware of the choices that shaped the interpretations we offer. Second, this is one of many cases in which annotation dilemmas mirror and highlight theoretical issues – in this case the fact that the processes which regulate ellipsis resolution very often do not yield unique outcomes. These issues arise again in interesting ways in our discussion of modality in sluicing (section 5.2).

Developing protocols for arriving at reasonable paraphrases was perhaps the most difficult design challenge we faced in the project. Many who use our materials will be struck by cases in which alternative paraphrases seem to be available and some will be skeptical of the apparent privileging of the paraphrases we ultimately settled on.[6] Those who are most skeptical about these aspects of our process are of course free to ignore the paraphrases we offer, while using whatever other aspects of our annotation-scheme they find useful. Our own view, though, is that this would be shortsighted. The paraphrases encode for each example, informally but accurately, the most salient reading perceived by annotators, often after considerable introspection and discussion. They imply no theoretical claims or commitments. What they provide is a hopefully useful set of empirical metrics against which proposals can be assessed. Successful theories of sluicing will yield for any sluice in our collection at least the interpretation corresponding to its paraphrase – by way of whatever assumptions or mechanisms seem right to their designers.

[6]Consider a particular example. The decision to begin the process by copying an antecedent into the ellipsis site and then modifying it brought about a welcome and significant increase in levels of agreement among annotators. But it also penalized paraphrases which are more distant in form from the antecedent. For that reason, as pointed out by a reviewer, the move probably led to an undercount of cases in which a paraphrase involving copular structures (so-called 'pseudo-sluices' or 'nonisomorphic' sluices, in the sense of van Craenenbroeck (2004, 2010a,b), Barros et al. (2014), Vicente (2019: 4.1)) would be appropriate. The decision to begin with a copying operation was prompted by two concerns. First, annotators in early pilots found simultaneous consideration of these kinds of alternatives as well as more syntactically isomorphic forms extremely taxing. They also proved difficult for us to adjudicate and analyze in our group meetings, since pseudo-sluice paraphrases often contain context-dependent expressions like discourse anaphora (*it* and *that*) and elisions (e.g., in the case of reduced clefts) which require their own, distinct annotation protocol. In that respect, they seemed unhelpful to investigators without deeper annotation. Non-isomorphic sluices are, nevertheless, not uncommon in our materials – see section 5.4 below.

4. INITIAL FINDINGS. Having described in broad terms how our dataset was constructed and how it can be queried, we turn to some empirical findings that emerge from it, focusing first on some very general characteristics of the data, turning then to a more particular focus on antecedents and the nature of the antecedent ellipsis relation.

4.1. GENERAL CHARACTERISTICS. An important theme in research on sluicing is the distinction between cases in which the WH-remnant has a counterpart in the antecedent context (a 'correlate') and cases in which it does not. The terms 'merger' and 'sprouting' (from Chung et al. 1995) are often used for the two kinds of cases. In (7a), there is a correlate for the WH-phrase (*some difference*) and it is therefore an instance of 'merger'. In (7b) there is no (overt) correlate.

(7) a. MERGER:
"It will make some difference, but I don't know how much," said A. Michael Lipper, president of Lipper Analytical Services Inc. [100549]

b. SPROUTING:
For the first time, Silver indicated that he was ready to vote on the plan, although he declined to say which way. [100447]

It is easy to identify the two types in our data – the CORRELATE tag applies exclusively to instances of merger; everything else is an instance of sprouting. And it is then surprising to observe that cases of sprouting outnumber cases of merger by a large margin – 65.5% to 34.5%. We call this observation 'surprising' because the relevant literature has tended to focus on cases of merger, although they represent, as it turns out, very much the minority case.

The high frequency of sprouting is explained in part by the enormous frequency of *why* as a WH-remnant. *Why*-sluices account for 53.8% of all instances of sprouting and 37.2% of sluices overall.[7] *Why* sluices are overwhelmingly, but not exclusively, of the sprouting type.

This is not the only surprise which emerges when we examine the distribution of semantic types (of the WH-remnant) in sluicing. The relevant findings are presented in Table 1. After remnants of type Reason (typically *why*), expressions of Degree (*how much, how tall, how often*) represent the second most frequent type – 22% of all of our data, followed by Entity expressions at 13.8% and Manner expressions at 7.2%. Once again we emphasize these figures because the literature on sluicing seems to have focused on the Entity type at the expense of other, more richly attested, kinds of cases. Consider three much cited publications on sluicing, for instance. In Ross 1969 7% of examples discussed are of the Degree type, 3% are Reason sluices, and 59% are entity sluices. For Chung et al. 1995, the proportions are 7% (Degree), 3% (Reason) and 79% (Entity), while for Merchant 2001 the proportions are 5% (Degree), 7% (Reason) and 60% (Entity).

Degree sluices in addition present a particularly rich set of puzzles, which have been little discussed, as far as we are aware. One of those puzzles (which initially emerged, again, as a quandary about how to annotate) is a subtle but widespread ambiguity, of the type seen in (8):

(8) Unisys said it would fire more employees, though it didn't say how many, and write off another $400 million against profits. [44148]

[7] The frequency of *why* sluices may be even higher, since the count given in the text excludes *why not* questions, which should probably not be analyzed in the same terms as sluicing (see Hofmann 2018 and the discussion in section 4.2 below). If such cases are included, then *why* sluices represent fully 53.8% of our total and 62.8% of instances of sprouting.

	SYNTACTIC POSITION		
SEMTYPE	EMBEDDED	ROOT	TOTAL
Reason	1642	110	1752
Degree	685	353	1038
Entity	335	315	650
Manner	290	45	335
Temporal	253	9	262
Locative	137	20	157
Classificatory	15	45	60
Other	47	399	446
Total	3404	1296	4700

Table 1: Distribution of annotated sluices by Semantic Type and Syntactic Status

On one reading of (8), the question under discussion is the absolute number of employees that the company might lay off. On an alternative reading, (8) raises a question about how much larger the number of employees to be laid off is than some pragmatically given point on the scale of expectation. This second interpretation might also be expressed by (9):

(9) Unisys said it would fire more employees, though it didn't say how many more, and write off another $400 million against profits.

The variables bound in the two interpretations are different and it is an interesting question what the source of those different variables might be in the antecedent context. It seems clear that a sustained investigation of the interaction between sluicing and constructions of comparison and degree would be revealing about both. Examples like (8) can be found by searching for the semantic type DEGREE and in addition searching for the tag REMNANT ELLIPSIS, which identifies cases in which an 'additional' ellipsis seems to apply within the WH-remnant, reducing, on one reading of (8), *how many more* to *how many*.

4.2. ANTECEDENTS. A fundamental issue in research on ellipsis has been the question of whether ellipses require overt linguistic antecedents and the subsidiary question of what kind of relation the antecedent relation is. Is it perhaps purely anaphoric, or are there parallelism conditions that must hold between the antecedent and the material to be elided? If there are such parallelism requirements, what form do they take? Our materials let us address these questions, for sluicing, in a precise and quantifiable way.

In typical cases, there is an antecedent and that antecedent is very local to the WH-remnant and precedes it. In the large majority of cases, the antecedent is in the immediately preceding sentence. But there are many atypical cases as well. There are, for instance, 115 cataphoric sluices, in which the sluice precedes the antecedent. Such cataphoric sluices are remarkably regular in form – in all but one instance, the antecedent occurs in the same sentence as the sluice and they are almost all of the form in (10):

(10) I don't know why, but [I said yes]. [144127]

Fifty five of these 115 examples have the exact string, 'I don't know why'.

Beyond these, there are 42 cases in which the sluice appears within its own antecedent (what we call "interpolated" sluices), as in (11):

(11) [A lot of people], I don't know for what reason, [are telling lies]. [57184]

In such examples the antecedent consists of two spans (*a lot of people* and *are telling lies* in (11)), which together form a clause but which are separated by the sequence of the WH-remnant and its embedding environment. In all 42 examples, that sequence (*I don't know for what reason* in (11)) is parenthetical. Sluicing is in general optional, at the cost of some small awkwardness, but for cases like (11) the cost of not eliding is unexpectedly severe:

(12) ??A lot of people, I don't know for what reason they are telling lies, are telling lies.

This may indicate that cases such as (11) do not involve ellipsis at all, but if that is the case then challenging questions arise about how WH-movement can have applied.

For the larger class of discontinuous antecedents (cases not involving interpolation of the remnant WH-phrase) questions also arise. (13) is typical.

(13) He turned toward [that part of the sky], which then [remained dark for a few seconds]. "It's hard to know how long [that part of the sky remained dark for]," he said.[125447]

In (13), the antecedent consists of the two spans *that part of the sky* and *remained dark for a few seconds*. There are 488 such cases. If the conventional wisdom is correct that antecedents are phrases rather than mere strings, then in the case of (13), the head of the appositive relative (*that part of the sky*) must be composed with the VP (*remained dark for a few reasons*) – by way of reconstruction or by way of a chain of anaphoric links (or both). Such cases are legion in our materials and would surely repay systematic investigation. A particularly interesting subclass of this type involves coordinate structures – cases (numerous) in which the antecedent is assembled from one piece which is external to the coordination and distributes over it and a second piece which consists of just one of the conjuncts. The two examples in (14) are representative.

(14) a. I get into the cupboards. Then for one year, I call them on a weekly basis and hold them accountable." In what way [do you hold them accountable]? [157514]

b. Messier was a distant second at 22 but has demanded a trade. He won't say why [Messier has demanded a trade], and he declined to play this year. [53758]

Such cases reveal that the antecedent relation, if there is such a relation, is no respecter of the integrity of coordinate structures. Cases like (14) can be found either by searching for the tag IGNORE or by searching for the string *coordination not interpreted*. In virtually every case, it is the rightmost conjunct which is shared with the ellipsis site.[8]

The fundamental question, though, is whether antecedents are required. As a matter of fact, 193 (4.1%) of our 4,700 examples lack antecedents. The examples in (15) are typical:

(15) a. Sam Kamvar, manager of the Childe Harold pub and restaurant, is proud of the framed

[8]Note incidentally that these are all cases in which our decision to use text spans for our annotations, rather than hierarchical syntactic or semantic representations, leaves open – entirely appropriately – the many analytical and theoretical questions that they raise.

poster that hangs on a brightly lit brick wall. "The Only Sign of Life in Dallas," it reads, above a highway sign: "Washington, D.C., 1304 Miles." Under that, it says, "Go Redskins." "I didn't ask *how much*," he said. "I bought it." [104061]

b. "I wanted to see everything that happens I wanted to hear everything that happens," said Charles Tomlin, whose 46-year-old son, Rick, an enforcement officer for the Federal Transportation Department, was killed in the blast. … "I had also wanted to know *why*. I wanted to look up and see McVeigh, a nice-looking man, and try to understand what drove him to kill this many people." [283235]

But 23 of these examples were judged to be of medium or low acceptability and three are cases in which an antecedent was almost certainly present in the conversational exchange but not reported in the article. The total of wellformed antecedentless examples then is actually 167 (3.5%).

Every fragment interrogative phrase in our corpus is initially categorized as a 'root sluice'. But whether or not all such examples involve ellipsis is far from clear. Among these examples, for instance, are five involving the conventionalized use of *how much* seen in (16) and in (15a).

(16) While the study consumed a $275,000 grant, the device it produced is a relative bargain. *How much*, installation included? "Oh," said Mehta, "I would suspect no more than $50 or so." [19606]

A much larger number (59 examples) involve a particular kind of fragment rhetorical question, discussed by Ginzburg & Sag (2000) and by Fernández et al. (2004). (17) is typical:

(17) Every day brings new evidence that the once-booming national economy is slowing. Even Alan Greenspan says so. But take a walk almost anywhere in Manhattan jostled by the leather-wearing, cell-phone-wielding, taxicab-grabbing, shopping-bag-swinging hordes and you have to wonder: *What slowdown*? Economists here are wondering, too. [29529]

Such expressions (all of the form *what* NP) communicate negative existential claims ('There is no slowdown' in (17)). They clearly deserve further study, but it is not clear that they share enough properties with sluicing that we should take the two to reflect the same grammatical mechanisms.[9]

A further large subgroup of 'missing antecedent' cases involves a modalized use of *why not*. In our materials, two clearly distinguishable uses of *why not* can be identified:

(18) a. FREE MODAL READING:
A: Should we go to the beach? B: Why not?

b. ANAPHORIC READING:
A: Frank doesn't believe in minimalism. B: Why not?

The anaphoric type in (18b) is elliptical in a standard sense, requiring an antecedent clause (there are 141 examples of this type and all have clausal antecedents). In fact its antecedent must be a 'negative clause' in exactly the sense of Klima (1964). Crucially their interpretation involves a

[9] All such examples in our dataset can be found by searching for the tag ECHOQ.

cancellation effect – although there is an expression of negation in the remnant and also in the antecedent, the elliptical question expresses a single negation.

The 'free modal' type seen in (18a) is very different. It does not require a linguistic antecedent, it is inherently modal in its interpretation, and (like (17)) is rigidly restricted to root contexts. There are 12 such examples among our antecedentless cases; those in (19) are typical.

(19) a. He learned English by listening to the radio. His first years were difficult. He lived in the basement of a building in Elmhurst, Queens, where Chinese immigrants paid $175 a month for beds separated by hanging blankets. He worked long hours in a laundry. Then he noticed some musicians in the subway. *Why not*? "I was so scared," Chen said. "I hesitated almost an hour. Then I counted to 100. Then I counted to 50. Then I finally opened the case." [95457]

b. Whelan said Ellison was an excellent student. He hadn't had a name athlete come through his door before. But Brown had read about Whelan in a fitness magazine and the gym was near Ellison's summer home. *Why not*? "I had heard and read that Pervis didn't work hard, but that was not the case with me," Whelan said. [241354]

Such cases are not easily assimilated to sluicing. Hofmann (2018) argues that the 'free modal' type in (19) is not elliptical and that the anaphoric type in (18b) involves not sluicing but rather a smaller ellipsis – one involving elision of the complement of a polarity head realized as *not*.

If such cases are also excluded, we are left with 91 examples out of 4700 – 1.9% – which are truly antecedentless.

There is more to be said, however. Of this group, just 34 are embedded, rather than root, sluices. This figure represents just 1% of embedded sluices. By contrast the 77 antecedentless root sluices represent 5.9% of root sluices. This confirms the speculation of Chung et al. 1995: 264-265 and Ginzburg & Sag 2000 that root sluices (or at any rate fragment WH-phrases) have greater freedom in licensing and in interpretation than their embedded counterparts.

An important subgroup of the remaining antecedentless cases is a class of *why*-sluices that we came to call 'Situational *why*' – as in (20):

(20) a. To read some accounts of his brief tenure at Georgia, one imagines a befuddled and confused Jim Harrick, huddled behind a locked door in his Stegeman Coliseum office, fighting back tears and wailing, "*Why, Lord, why*?" [103083]

b. McCann said Jonesboro, a town of 51,000 on the Mississippi River, drew closer together as a result of the March 24, 1998, shooting rampage during which four junior high school girls and one teacher died, "but there's a lot of anger," he said. "The biggest problem for people I've talked to is there's never been an answer to *why*. I think if anyone could ever answer *why*, it would help a lot, but I don't think that's going to happen." [109969]

This use of *why*, in both root and embedded settings, expresses bewilderment about why dreadful things happen (never neutral things or happy things) and the requests for enlightenment are often addressed to the deity. If such uses of *why* are also best regarded as conventionalized and perhaps nonelliptical, the number of truly antecedentless examples goes yet lower: 0.3% (10 of 3404) of embedded sluices and 3.2% (42 of 1296) of apparent root sluices.

Of course it is important now to ask if similar patterns would emerge for other genres – in

conversational exchanges in particular. But for the moment, the conclusion seems to be that, to comport with observation, our theories need to guarantee that some 99.7% of embedded sluices have antecedents. But they also need to provide an understanding of why root and embedded sluices are different in this respect – root sluices tolerate the absence of antecedents measurably more frequently than embedded sluices do.

5. The dimensions of mismatch. But if in the vast majority of cases sluices have antecedents, how closely must that antecedent resemble the elided clause and in what ways? Here we map the principal patterns of difference attested in our dataset. We were surprised by the range of possibilities that emerged.

5.1. Tense. There are 129 instances of tense mismatch, an annotation used when there is a tense form implied by the interpretation of the sluice which does not seem to match that found in the antecedent. In 36 instances, the mismatch can be understood as primarily syntactic. In 20 of these cases, the antecedent lacks a syntactic expression of tense, as in (21), where the antecedent is a gerund. In the 16 remaining cases, the antecedent includes an overt tense expression, but one that is syntactically and semantically distinct from that implied by the meaning of the sluice. In 8 of these, this is a modal auxiliary which the paraphrase lacks. One such example is (22), where the epistemic modal *must* is not retained (but the past perspective signaled by *have* is). In the 8 remaining cases, the antecedent and paraphrase disagree in finiteness,

(21) She remembered [Ronnie spending six months in some kind of "school for boys"] when he was a youth but she doesn't know why [Ronnie SPENT six months in some kind of school for boys]. [125278]

(22) From what I can make out, [it must have been written] sometime during the Vietnam War, but I don't know by whom [it WAS written]. [72508]

The remaining 93 instances all involve mismatch between finite tense morphemes. For 20 of these, the antecedent is in a quotation and the sluice is outside the quotation, meaning that the same event is viewed from different temporal perspectives; a representative example is in (23).

(23) You heard the question, "Why [are people watching this]?" For once, I didn't care why [people WERE watching this]. [48694]

The final 73 all devolve to matters of how tense morphemes behave in English embedded clauses. In some cases, the issue is simply sequence of tense, as in (24), where the same clause is embedded first under the past tense *told* and then, in the sluice, under the present tense *know*.

(24) "I told him [I'd support him in his efforts and be an investor]," Kemper said Monday. "I don't know to what extent [I WILL support him in his efforts and be an investor] yet, because I've just decided to do it this morning." [15852]

(25) is similar: an event is introduced in a historical present narrative, and is then commented on from a temporal perspective after the event has transpired, requiring a past tense.

(25) Everyone exhibits it of course. People misplace their keys. [They enter a room] only to realize they don't know why [they enter-ED THAT ROOM]. [211474]

Such cases seem to pose problems for any morphosyntactic identity condition for sluicing, since past and present tense are distinct morphemes (as are the modal auxiliaries). In addition, the instances with the tenseless or modalized antecedents are rather surprising under a view where one simply copies or retrieves a clause level antecedent.

One might view these observations as arguments in favor of semantic identity theories, since many mismatches could be understood as analogs of the kinds of 'vehicle change' observed for anaphoric reference under other kinds of ellipsis. For instance, under a referential theory of tense, a past morpheme and a present morpheme can be denotationally equivalent, just as *I* and *he* can corefer in the right environments. However not all mismatches involving temporal interpretation are of this anaphoric character. In several *for how long* sluices in the database, a present tense antecedent is paired with a vaguely modal or future interpretation in the paraphrase, as in (26) below.

(26) [Rob and Mike both still fish], but they don't know for how long [they {will, might, could, . . .} fish]. [138058]

As we will see shortly, there are also many instances where an overt intensional operator in the antecedent is replaced by a similar, or related, modal paraphrase in the ellipsis site. In (26) and in examples like it, the antecedent has no intensional operator, but annotators nonetheless included a modal in the paraphrase, because the simple present does not accurately reflect the interpretation of the elided clause. If this interpretation is correct, sluicing must tolerate semantic mismatches in tense alongside the apparently syntactic mismatches we began with.

5.2. MODALS. In some 394 of our annotations, the paraphrase of elided material contains a modal not present in the antecedent. Such cases can be found by searching for the symbol MODAL, which indicates the presence in the ellipsis site of a modal of some flavor. Very often, it is difficult to identify the implied modal with any particular English modal verb.

Some of these annotations, however, are open to the charge that they reflect nothing more than choices forced by our annotation guidelines – by the policy, in particular, which favors the smallest antecedent consistent with the observed interpretation (see section 3.2 above). Consider the examples in (27). In each case (i) is the paraphrase offered in our materials, while (ii) is an alternative which assumes a larger antecedent. In each case also, (i) postulates a modal mismatch, while (ii) does not.

(27) a. In his state of the union message last week, Clinton said [he] favored [raising the minimum wage] but did not say by how much. [15642]
(i) [he MODAL raise the minimum wage]
(ii) [he **favored** raising the minimum wage].

b. Arizona officials concede [this year's exports] are likely [to slow], but they do not know by how much. [54079]
(i) [this year 's exports MODAL slow]
(ii) [this year's exports **are likely to** slow]

Although both alternatives seem reasonable, the policy favoring smaller antecedents forces (i). The interesting property highlighted by the annotation dilemma (again) is that in such cases there is in the antecedent context an intensional expression which has the smaller of the two potential antecedents in its scope (*favored* in (27a), *likely* in (27b)). The modal base characteristic of the

ellipsis site is anaphoric to the intensional context established by that embedding predicate. But there is a near equivalent alternative annotation which assumes a larger antecedent and repetition of the embedding verb in the elided clause. The judgment calls concerning interpretation in such cases are very delicate indeed (compare example (6) above) and since we know of no principled way to decide which alternative is more accurate, we set such cases aside here – while recognizing the important analytical questions that they raise.

Even when such cases are set aside, however, many examples remain for which no evident alternative to the postulation of modal mismatch is available. Those in (28) are typical.

(28) a. Now comes the hard part. With all this banter about cyberspace, is it worth it [to get your student on line]? If so, how [MODAL YOU get your student on line]? With what company? [MODAL YOU get your student on line] [23721]

b. Oz Chairman Robert Kory vowed [to push ahead] but would not say how [Oz Chairman Robert Kory MODAL push ahead]. [30594]

In all of these cases, the antecedent is nonfinite. Similar effects are observed with imperative antecedents:

(29) a. "Turn at the next corner," my wife said. I didn't ask why [I MODAL turn at the next corner]. [142535]

b. When it comes to mail-order purchases, always use a credit card. [Never pay cash]. And you know why [YOU MODAL never pay cash]? [47922]

Such cases, however, by no means exhaust the space of possible modal mismatches, as we see in the examples of (30). In these and similar cases, tense and modality are fully specified in the antecedent, but not in ways which match that required by the sense of the sluice:

(30) a. A minute later, though, he denounced the release anyway, saying [it should have happened] three or four years ago. How [it MODAL have happened], he didn't reveal. [F50]

b. And then you get (Pam) Shriver finding a bald man standing at the fence with a giant-sized tennis ball in his hand, asking, " [Would you sign my ball]?" So, she darted her red eyes from his head to the ball: "Which one [MODAL I sign]?" [205304]

c. Some anti-AOL types are using cyberspace bulletin boards to try to rally users to oppose it, as well. If [enough users opt out] – I don't know how many [users MODAL opt out]; the number's not disclosed – the settlement will be voided. [71495]

The empirical territory here is again fascinating and delicate, since the exact sense of the modal is often underdetermined. Some conclusions are clear however. (30a) allows or demands an ability/possibility modal. In (30b) the required modal seems closest to *should*. The interpretation of the sluice in (30c) requires a necessity modal, but that aspect of its meaning seems to have its source in the modal semantics of the conditional, rather than in any element of the antecedent clause itself. In fact in none of these cases is the sense of the elided modal determined by any item by item isomorphism with a corresponding element in the antecedent. That conclusion is reinforced by cases like (31), in which the antecedent is subclausal – a nonverbal small clause which lacks any expression of tense or modality. Yet in both cases the sense of the elided clause implies the presence of a possibility modal.

(31) a. Among the proposals are [new power plants in the region], although the report does not specify where [THOSE new power plants MODAL BE in the region]. [143606]
b. The rediscovered inspiration for "Ted Williams" brought [the burglar back], but his creator can't say for how long [the burglar MODAL BE back]. [124186]

5.3. POLARITY. Somewhat surprisingly, we encountered 28 cases in which the antecedent clause and the elided clause differ in polarity, as in (32):

(32) "[Coach O'Leary doesn't do things] without letting you know why [Coach O'Leary DID THOSE things]," Hamilton said. [99992]

Note that it is apparently harder to 'add' negation between antecedent and ellipsis site than to 'subtract' it – 23 cases, like (32), involve a negative antecedent and a positive ellipsis site, while only five cases (11174, 22987, 915941, 99105, F74) show the opposite direction of reversal.

An important question that now arises is why cases of polarity mismatch seem to be so rare. Examining a set of cases uncovered earlier by our project, Margaret Kroll (2016, 2019) proposes that polarity reversals under sluicing are only possible if the discourse context containing an apparent antecedent a is such as to render the proposition $\neg a$ salient in, and entailed by, the local context. Polarity reversals under sluicing should be pragmatically licensed, then, if the uttering of a negative proposition triggers a local context update containing the positive counterpart of that proposition; or vice versa. Such contexts are not routine and mismatches should then be finely sensitive to properties of the local discourse context. This seems to be the case. We survey five such context types.

The first, unsurprisingly, involves NEG-raising triggers, as illustrated in (33).

(33) "I don't think [Steve Jobs will let it be a boring MacWorld]," Reynolds said. "We just don't know how [Steve Jobs will let it be NOT a boring MacWorld]." [111174]

I don't think P is essentially an alternative way to make salient and given the proposition *not P*. This is presumably why NEG-raising so facilitates polarity reversal under sluicing. A similar effect holds for cases involving disjunction or embedding predicates like *doubt* and *remember*:

(34) a. Angela J. Campbell, an attorney for opponents to the deal, told the Globe that McCain's letter likely "tipped" the scales in favor of the decision. "Senator McCain said, '[Do it] by December 15 or explain why [YOU DID NOT do it],' and the commission jumped to it and did it that very day," Campbell told the Globe. [22987]
b. "It raises real doubts as to whether [Iraq is ready to give full, final and complete disclosure]" of its weapons programs, said the UK's Weston. "One has to ask oneself why [Iraq is NOT ready to give full, final and complete disclosure]?" he said earlier in his briefing with reporters. [99105]
c. But he was handed a small Belgian pistol, and he had little choice but to stay and help, harassing Japanese patrols by night and trying to defend a small patch of land against a communist takeover. "I don't know why [I WAS NOT scared], but I really cannot remember being scared," he said. "It all seemed like great fun." [91594]

Kroll (2019) argues that in such cases, pragmatic calculations make the negation of the antecedent locally given and salient, either because of the typical pragmatics of disjunction (Karttunen 1974)

or a constellation of defeasible background assumptions about memory.

Building on Kroll's logic, we can see that a similar state of affairs holds for example (32). In this case, *without* serves to support the inference (locally, under negation) that Coach O'Leary does things, linking such events with explaining events by O'Leary. Polarity reversal is thus an expected option here. Cases like (32) were discovered by Masaya Yoshida (2010) and discussed also by Lasnik & Funakoshi (2018: 66–68). Yoshida argues that the crucial factor permitting (32) is the fact that *without* phrases adjoin to VP. But to our knowledge, these kinds of polarity reversals are tolerated only when the relevant adjunct is headed by *without*, suggesting that its particular semantic and pragmatic properties are the crucial factor.

Finally, in the fifth kind of discourse context, a sequence of (negative) partial answers to some superordinate question under discussion (a QUD in the technical sense introduced by Roberts (2012)) licenses the overt raising of that QUD. In such contexts, the positive counterpart of A's negative assertion (though never uttered) forms the basis for an application of sluicing because it is presupposed by the superordinate QUD and is therefore given in the required sense. The examples in (35) provide instances:

(35) a. That kind of money is now the biggest challenge to America's democracy. And yet since both sides were flinging it around, money can't be said to have determined the outcome. What [can be said to have determined the outcome] then? [41116]

b. "He came back and asked me why I had put the two ballets together," she says. He had wanted it, she reminded Balanchine. "But that doesn't mean all the time," Balanchine chided her. How many times [does it mean] then? " Well, four," he said at last. [152316]

One of the characteristics of such contexts is that the discourse particle *then* often accompanies the WH-question (independent of sluicing); in many cases, it is close to obligatory. Biezma (2014) has studied such uses of *then* and her account seems to extend to cases like (35). Her proposal is that *then* is subject to a felicity condition which demands that its antecedent and consequent reflect a causal explanatory claim. According to this view, the assertion of a sequence of one or more negative propositions $\neg \mathrm{P}(c)$ – incomplete answers to a QUD, explicit or implicit – licenses an inference of the form $\exists x \mathrm{P}(x)$, prompting the WH-question which seeks a complete answer and optionally licensing sluicing in virtue of its givenness and its relation to the QUD. The information gain from the discourse move introducing the negative assertion is part of the explanatory chain linking the assertion event with the question event in the unfolding of the discourse.

There is an additional type of polarity reversal under sluicing, though, which is productive and does not seem to require careful contextual staging. In these cases *how* is the remnant WH-phrase and the question embedding predicate is *know* or *learn* or *see*. There are 13 examples of this kind in our corpus, and they therefore represent the single largest group of polarity reversing examples so far uncovered; the examples in (36) are typical.

(36) a. "They obviously haven't tried any cases in a long time, and obviously don't know how, [they MODAL try cases] but this is cross-examination." [123640]

b. Republicans cannot compete with Clinton at this level of the game. They don't know how [they MODAL compete with Clinton] [132033]

In all such cases, negation in the antecedent context has no counterpart within the elided clause. It can hardly be an accident, in addition, that in 12 of the 13 cases the clause in which the sluice

is embedded is negated (as in (36)).

It remains unclear, at present, whether or not such cases fall under Kroll's proposal (why would the assertion of $\neg p$ make p salient and locally entailed underneath *don't know how* in (36)?). This is just one of many questions which can now be taken up, but the phenomenon itself (polarity reversal under sluicing) is clearly a robust one.[10]

5.4. New words. Beyond the specific mismatches discussed so far, some 160 examples in our dataset are marked with a more general New Words tag, which indicates that the paraphrase contains lexical material not found in the antecedent. A persistent idea in research on ellipsis has been that antecedent and ellipsis site must be parallel in being composed of the same lexical resources assembled in the same way (Ross 1967: 5.135, p. 348, Rooth 1992, Fiengo & May 1994, Chung et al. 1995, Heim 1997, Chung 2006, 2013, Rudin 2019). The New Words tag is designed to help assess that claim.

Setting aside cases where use of the tag might reflect only the demands of our own guidelines (see section 3.2) and cases annotators judged as illformed, there are 71 clear cases. Three subtypes can be identified – 46 involve copular clauses in the ellipsis site, 17 involve existential interpretations in the ellipsis site, and 17 involve stranded prepositions in the ellipsis site which have no counterpart in the antecedent.

Turning to the prepositional cases first, descriptively speaking the role of the 'missing' prepositions is to enforce idiosyncratic grammatical restrictions characteristic of particular adjunct types, *in what decade*, *on what night*, *piece from 1991* , or *at which firm* in (37).[11]

(37) a. He says [America was once a better place] and that he knows it because he was there. What decade [was America a better place IN]? [138872]

b. Then you see where [they're going to place it]: What night [are they going to place it ON]? [138731]

c. "The first thing he said was so interesting that [he thought it was a period piece]," Scardino recalled. "I said 'What period [do you think it is a piece FROM]?' He said, 'Nineteen ninety-one." ' [195676]

d. Decker was weaned in the world of investing by his father, who had also been a mutual fund manager. (Decker won't say which firm [his father had been a mutual fund manager AT]). [89932]

Beyond these cases, a large fraction involved clause building functional vocabulary. One set of these begin from a nominal antecedent, which is added to in the paraphrase to construct a wellformed clause. In most cases, annotators proposed a copular clause, in the process creating the copular 'nonisomorphic' sluices of recent discussions – see van Craenenbroeck 2004, 2010a,b, Barros et al. 2014, Vicente 2019: 4.1:

(38) a. Bradley said that he has not shut the door to [a presidential race], though he would not say when [that presidential race MODAL BE]. [176498]

[10]It is tempting, for English, to treat cases like (36) as deriving from an infinitival source (*don't know how to VP*). That may well be a reasonable move in syntactic terms, but it does not, in and of itself, void the inference that such examples involve polarity reversal. An anonymous reviewer, in addition, points out that reversals like those in (36) are also found in languages which lack infinitives. If that is so, then appeal to a possible infinitival source will not yield a full understanding of the phenomenon. Other, or additional, factors must be in play.

[11]Novel material in the paraphrase is indicated by upper case.

b. The doctors anticipate [a full recovery] for me, but they really don't know when [that recovery MODAL BE]. [76117]

In 17 cases, however, the paraphrase was an existential construction:

(39) a. [A cut] appears almost certain this year; the question is how soon [THERE MODAL BE a cut], and by how much [THERE MODAL BE a cut]. [15811]
b. Even the most conservative voices in the state seem resigned to the prospect of [a long costly court battle]. To what end [MODAL THERE BE a long costly court battle]? [135056]

For an additional 23 cases, the antecedent was not simply a nominal, but an embedded small clause, which was again enlarged to a copular clause in the paraphrase:

(40) a. The bodies were discovered just before 1 a.m. when an employee of the shop happened to drive by, noticed [lights still on] almost three hours after closing time and went inside to see why [the lights WERE still on]. [72082]
b. I don't know when [THERE MODAL BE a couple of major league teams in Japan, one in Seoul, and one in Hawaii], but I can see [a couple of major league teams in Japan, one in Seoul and one in Hawaii], as the stopover on the way. [72698]

While there are many intricate subquestions that arise for each of these cases of novel lexical material in sluicing contexts, as a group they pose serious challenges for all versions of the lexical parallelism constraints that have been proposed to date. We address many of these questions in Anand et al. 2020.

5.5. MISSING MISMATCHES. Our focus so far has been on mismatches in interpretation and form between the elided clause of a sluicing construction and its apparent antecedent. It is just as important, however, that we document what has not been observed. In particular, we found no cases which challenge the claim that argument structure congruence must hold between the elided clause and its antecedent. We have observed no antecedent ellipsis site pairings in which one was active and the other passive, in which one was transitive and the other inchoative or in which one had a double object structure while the other had a nominal prepositional phrase structure. Of course it does not follow from their nonappearance in our corpus that such mismatches are impossible – they may simply be very rare. But it is important that we report the fact. In addition, the contrast with VP ellipsis here is striking. Early work on VP ellipsis assumed that active passive mismatches were impossible. But naturally occurring examples were noted as early as Sag 1976, and, when serious corpus work began, numerous examples very quickly came to light (Dalrymple et al. 1991, Hardt 1993, 1997, Kehler 2002, Merchant 2013). VP ellipsis seems to be less frequent than sluicing. If voice mismatches were possible for sluicing, it is hard to see why naturally occurring examples should be so much more elusive than they are for VP ellipsis.

6. CONCLUSION. In creating and making available this dataset, our principal aim has been to provide researchers with a new kind of tool by means of which the mysteries of ellipsis can be further probed. We believe that it will help resolve existing questions and we hope that it will provoke the asking of new ones. We have tried to show in this overview that it has that potential. Already, it has brought to light kinds of mismatch between ellipses and their apparent antecedents which have either gone unnoticed or been little examined (Kroll & Rudin 2018, Kroll 2019, Rudin

2019, Hardt & Rudin 2019). Such findings radically and usefully narrow the hypothesis space for theories of ellipsis licensing and resolution, as we show in a companion paper (Anand et al. 2020).

At a more programmatic level, we have shown that it is possible to create a large annotated dataset in a theoretically challenging domain at a high level of reliability and sophistication. There is every reason to believe that the same could be achieved for other domains – equally complex and equally important – and that such efforts would likewise yield valuable new puzzles, challenges, and insights.

References.

Anand, Pranav; Daniel Hardt; and James McCloskey. 2020. The domain of matching in sluicing. Manuscript, University of California Santa Cruz.

Anand, Pranav, and Jim McCloskey. 2015. Annotating the implicit content of sluices. *Proceedings of the Ninth Linguistic Annotation Workshop*, LAW IX '10, 178–187. Denver, Colorado, usa: Association for Computational Linguistics.
url: https://www.aclweb.org/anthology/W15-1600/.

Barros, Matt; Patrick Elliott; and Gary Thoms. 2014. There is no island repair. Available on LingBuzz. url: https://ling.auf.net/lingbuzz/002100.

Biezma, María. 2014. The grammar of discourse: The case of *then*. *SALT XXIV, Proceedings of Semantics and Linguistic Theory 24*, ed. by Todd Snider, Sarah D'Antonio, and Mia Weigand, 373–394. lsa and clc Publications.

Chung, Sandra. 2006. Sluicing and the lexicon: The point of no return. *BLS 31, Proceedings of the Thirty-First Annual Meeting of the Berkeley Linguistics Society*, ed. by Rebecca Cover and Yuni Kim, 73–91. Berkeley California: Department of Linguistics, uc Berkeley.

Chung, Sandra. 2013. Syntactic identity in sluicing: How much and why. *Linguistic Inquiry* 44.1–44.

Chung, Sandra; William Ladusaw; and James McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics* 3.239–282.

van Craenenbroeck, Jeroen. 2004. Sluicing in Dutch dialects. Doctoral dissertation, Leiden University.

van Craenenbroeck, Jeroen. 2010a. Invisible last resort: A note on clefts as the underlying source for sluicing. *Lingua* 120.1714–1726.

van Craenenbroeck, Jeroen. 2010b. *The syntax of ellipsis: Evidence from Dutch dialects*. Oxford Studies in Comparative Syntax. Oxford and New York: Oxford University Press.

Dalrymple, Mary; Stuart M. Schieber; and Fernanda C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy* 14.399–452.

Fernández, Raquel; Jonathan Ginzburg; and Shalom Lappin. 2004. Classifying ellipsis in dialogue: A machine learning approach. *Proceedings of the 20th international conference on computational linguistics*, Association for Computational Linguistics, 240.

Fiengo, Robert, and Robert May. 1994. *Indices and identity*. Cambridge, Massachusetts: mit Press.

Ginzburg, Jonathan, and Ivan Sag. 2000. *Interrogative investigations: The form, meaning and use of English interrogatives*. Stanford, California: csli Publications.

Graff, David; Junbo Kong; Ke Chen; and Kazuaki Maeda. 2005. English gigaword second edition ldc2007t07. Tech. rep., Linguistic Data Consortium, Philadelphia.

HARDT, DANIEL. 1993. Verb phrase ellipsis: Form, meaning and processing. Doctoral dissertation, University of Pennsylvania.

HARDT, DANIEL. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 32.525–541.

HARDT, DANIEL, and DENIZ RUDIN. 2019. Sluicing and modal mismatches. presented at *Sluicing@50*, University of Chicago, April 12th 2019.

HEIM, IRENE. 1997. Predicates or formulas? Evidence from ellipsis. *SALT VII, Proceedings from Semantics and Linguistic Theory VII*, ed. by Aaron Lawson, 197–221. Ithaca, New York: Cornell University, CLC Publications. DOI: https://doi.org/10.3765/salt.v7i0.2793.

HOFMANN, LISA. 2018. Why not: Polarity ellipsis and negative concord. Manuscript, Department of Linguistics, University of California Santa Cruz.

KARTTUNEN, LAURI. 1974. Presuppositions and linguistic context. *Theoretical Linguistics* 10.181–194.

KEHLER, ANDREW. 2002. *Coherence in discourse*. Stanford, California: CSLI Publications.

KLEIN, DAN, and CHRISTOPHER C. MANNING. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430. Sapporo, Japan: Association for Computational Linguistics. Online: https://www.aclweb.org/anthology/P03-1000.

KLIMA, EDWARD S. 1964. Negation in English. *The structure of language: Readings in the philosophy of language*, ed. by Jerry A. Fodor and Jerrold J. Katz, 246–323. Prentice Hall.

KROLL, MARGARET. 2016. Polarity reversals under sluicing. *Proceedings of Sinn und Bedeutung 21*, Edinburgh: School of Philosophy, Psychology and Language Sciences, University of Edinburgh. DOI: https://semanticsarchive.net/Archive/DRjNjViN/index.html.

KROLL, MARGARET. 2019. Polarity reversals under sluicing. *Semantics and Pragmatics* 12, DOI: http://dx.doi.org/10.3765/sp.12.18.

KROLL, MARGARET, and DENIZ RUDIN. 2018. Identity and interpretation: Syntactic and pragmatic constraints on the acceptability of sluicing. *NELS 47: Proceedings of the Forty Seventh Annual Meeting of the North East Linguistic Society*, ed. by Andrew Lamont and Katerina Tetzloff, . Volume Two, 177–190. Amherst, Massachusetts: GLSA.

LASNIK, HOWARD, and KENSHI FUNAKOSHI. 2018. Ellipsis in transformational grammar. *The Oxford Handbook of Ellipsis*, ed. by Jeroen van Craenenbroeck and Tanja Temmermann, Oxford Handbooks, 46–74. Oxford and New York: Oxford University Press.

MERCHANT, JASON. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford and New York: Oxford University Press.

MERCHANT, JASON. 2013. Voice and ellipsis. *Linguistic Inquiry* 44.77–108.

ROBERTS, CRAIGE. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5.1–69.

ROHDE, D. L. T. 2005. Tgrep2 user manual, v. 1.15.
Retrieved from https://tedlab.mit.edu/~dr/Tgrep2/.

ROMERO, MARIBEL. 1998. Focus and reconstruction effects in WH-phrases. Doctoral dissertation, University of Massachusetts, Amherst.

ROOTH, MATS. 1992. Ellipsis redundancy and reduction redundancy. *Proceedings of the Stuttgart Ellipsis Workshop*, ed. by Steve Berman and Arild Hestvik, *Arbeitspapiere des Sonderforschungsbereichs*, . 340, 1–26. Heidelberg: Universität Stuttgart.

ROSS, JOHN R. 1967. Constraints on variables in syntax. Doctoral dissertation, MIT, Published as *Infinite Syntax!* Norwood: New Jersey: Ablex (1986).

Ross, John R. 1969. Guess who? *CLS 5: Papers from the fifth regional meeting of the Chicago Linguistic Society*, ed. by Robert Binnick, Alice Davison, Georgia Green, and Jerry Morgan, 252–286. Chicago Illinois: Chicago Linguistic Society.

Rudin, Deniz. 2019. Head-based syntactic identity in sluicing. *Linguistic Inquiry* 50.253–283.

Sag, Ivan. 1976. Deletion and logical form. Doctoral dissertation, mit.

Stenetorp, Pontus; Sampo Pyysalo; Goran Topić; Tomoko Ohta; Sophia Ananiadou; and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. *Proceedings of the demonstrations session at EACL 2012*. Project Website: http://brat.nlplab.org/about.html.

Vicente, Luis. 2019. Sluicing and its subtypes. *The Oxford handbook of ellipsis*, ed. by Jeroen van Craenenbroeck and Tanja Temmerman, Oxford Handbooks in Linguistics, chap. 20, 479–503. Oxford and New York: Oxford University Press.

Yoshida, Masaya. 2010. Antecedent-contained sluicing. *Linguistic Inquiry* 41.348–356.