1 Introduction

There are few phenomena in natural language more liable to be taken for granted than ellipsis. It feels natural to leave redundant content unexpressed; but from the comprehender's perspective, elision should hardly be 'natural.' And yet, it is. In (1a), the verb phrase is missing, while in (1b), it is repeated.

(1) a. Although I didn't really have to avoid meat for a year, I did.
b. Although I didn't really have to avoid meat for a year, I avoided meat for a year.

But rather than being helpful, this repetition strikes a sour note. Remarkably, we prefer the burden of working out what the speaker might have meant to the burden of listening to a repetition. Why should that be?

It may well be premature to ask such why-questions, but there are many preliminary problems to be tackled which, while challenging, seem more tractable—what the mechanisms of ellipsis are and what the range of crosslinguistic variation is. With these questions, every branch of the language sciences (theoretical linguistics, psycholinguistics, computational linguistics) must grapple. The first goal of this project is to build and make freely available to those who engage such questions, a new kind of resource, one designed to be valuable to theoreticians of every persuasion and also to those with more practical goals. The resource will be a large corpus of naturally occuring data, recording ellipses in their discourse context, annotated at a level of sophistication that will allow the hardest and most interesting questions to be investigated. It will be web-based, freely accessible to researchers of all kinds, and searchable by way of an intuitive and well-designed user-interface.

Many different species of ellipsis have been documented, and to keep things manageable, while at the same time constructing something useful, we will focus on the type known as *sluicing*:

(2) They're arriving tomorrow, but we don't know exactly when.

In sluicing all but the interrogative phrase of a content question is elided. We choose sluicing as our initial target because it is widely attested across languages (making it a good base for an extension beyond English), because it is well studied (which means that we have the makings of a rich annotation system), and because it interacts in interesting ways with many other aspects of linguistic form and interpretation—question-hood, 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.

As the work of building, testing, and presenting the resource proceeds, the PI's will make use of the emerging patterns to investigate a number of issues that have been central in theoretical debate. Our hope is that the size and richness of the database will make possible more definitive kinds of conclusions (empirical and analytical) than have been possible to date. Papers and presentations emerging from the project will be directed equally to theoretical linguists and to computational linguists.

2 The Research Context

First we fix some terminology. Consider the examples in (3) and (4).

(3) a. They've made an offer to [one of the candidates], but I'm not sure *which one*.
b. She bought [some] properties; *how many*, we're not entirely sure.

(4) a. They were firing, but *at what* was unclear.
b. He finished on time, but *with whose help*?

In interpreting sluices, comprehenders must 'resolve' the ellipses, filling out ellipsis sites based on contextually-supplied content and thereby interpreting the WH-phrases as full content questions. The literature recognizes three dimensions along which sluices may vary. Consider (3) and (4). In (3), the context contains a phrase which corresponds to the WH-phrase of the sluice—*one of the candidates* in (3a) for example. Here we will follow Chung et al. (1995) in using the term *inner antecedent* for such phrases and abbreviate it IA (the term *correlate* is also used). There are also cases in which the context contains no IA, as in (4). We will follow custom and use the term *sprouting* for such cases We will use the term *merger* for cases like (3) which lack an IA. Whether or not the distinction between merger cases and sprouting cases is more than terminological has been a major point of contention.

A second dimension of variation is the dialogic source of the context. We have so far considered only 'same-speaker' sluices, but 'cross-speaker' sluicing, which borrows from another's utterance, is common:

(5) A: I got a new laptop as a Christmas present.
B: Who from?

Finally, most of the cases we have seen are embedded, and embedded sluices can be complements, ((3a)), subjects ((4a)), or topicalized ((3b)). In addition, sluices may also be free-standing 'root sluices', as in (5) above. Although this is also cross-speaker, root sluices need not be cross-speaker, as is clear from (4b).

These then are the dimensions of the resolution problem for sluicing: The right analysis must explain how apparently free-standing interrogative phrases can somehow draw content from the context of use to be interpreted as if they were clauses. That content can be provided by the speaker of the sluice or another speaker. Some sluices occur in root contexts and some as complements; some are subjects and some are topicalized. Finally, there may or may not be an IA for the WH-phrase. Such variation prompts some natural questions: Are all types of sluice subject to the same constraints? Are they resolved by the same mechanisms? Existing answers to these questions are incomplete and disputed, as we will see. If we turn to the theoretical landscape, the central questions are the following:

QUESTION 1: What, if anything, is the content of the ellipsis site?
QUESTION 2: What is the relation that holds between the recovered content and the ellipsis site?

Three types of answers have been offered to these questions:

TYPE 1: The ellipsis site has no internal structure and its reference is resolved in the same way as for any other anaphoric element (Hardt 1993, 1999, Darymple et al. 1991, Schieber et al. 1999, Ginzburg & Sag 2000, Culicover & Jackendoff 2005, Barker 2013). On this view, the answer to QUESTION 2 is that the relation between the sluice and the recovered content is fundamentally anaphoric.

TYPE 2: The ellipsis site is empty before spell out, and is filled in later by re-using (recycling, copying) an already built syntactic structure, including interpretation, from the discourse (Williams 1977, Fiengo & May 1994, Lappin 1999, Chung et al. 1995, 2011). What is copied is crucially a fully articulated syntactic structure with its full interpretation (this answers]QUESTION 2). The answer offered to QUESTION 1 is that the ellipsis site acquires a syntactic structure at some non-surface level of representation—one relevant for the computation of meaning.

TYPE 3: The ellipsis site has ordinary internal syntactic structure which is silenceable because it is sufficiently similar to some antecedent XP elsewhere in the discourse (Ross 1969, Sag 1976, Hankamer 1979, Lasnik 2001, Merchant 2001, 2004, Craenenbroeck 2010 and many others). The answer offered to

QUESTION 2 is that the relation between the ellipsis site and the recovered content is similarity or parallelism.

The first of these answers is surely the most parsimonious, as it postulates nothing beyond an independently necessary theory of anaphoricity. Type 2 and Type 3 approaches, in contrast, are burdened with more ontological baggage. First, they both assume articulated syntactic structure within the ellipsis site at some level of representation (there is 'syntax in the silence', to use Merchant's term). Relatedly, they assume (unlike at least some Type 1 approaches) that sluicing requires an identifiable linguistic antecedent accessible in the discourse—a syntactic constituent to be re-used in the ellipsis site (Type 2) or to act as antecedent for the elision (Type 3).[1] Finally, they assume some additional mechanism in answering QUESTION 2—LF copying for Type 2 approaches, or deletion in phonology for Type 3 approaches. In turn, these operations must somehow allow syntactic or morphophonological operations to be sensitive to aspects of discourse-structure in a way that seems at odds with reasonable assumptions about the privileges of access allowed to these modules.

All of these approaches have their advocates[2] today. However, Type 3 approaches represent very much the consensus in those research traditions that derive from Principles and Parameters theory (Romero 1997, 1998, Lasnik 1999 and especially Merchant 2001). In this perspective, the central challenge is to say what it means for a phrase to be 'similar enough' to some antecedent phrase to be reducible to silence.

While early accounts (Ross (1969) for example) assumed simply that deletion is possible only under morphosyntactic identity with surface syntax, it quickly became clear that this condition is too stringent. However, with the emergence of Logical Form (LF), a level distinct from the surface form and readied for semantic interpretation, identity could be required of LF-syntax rather than of surface syntax. The representational abstraction implied by this shift removed many of the obvious difficulties faced by earlier syntactic treatments (Williams 1977, Sag 1976, Fiengo & May 1994 among others). In turn, the development of sophisticated semantic treatments of VP ellipsis (Darymple et al. 1991, Rooth 1992, Hardt 1993 for example), made possible a perspective in which deletion was licensed on purely semantic grounds. This was the synthesis argued for by Merchant (1999, 2001)— a Type 3 analysis of sluicing in which the antecedent and elided clause must entail each other, *modulo* focused material. In the decade and a half since this consensus emerged, a great deal of work has been devoted to probing and extending the proposal. Two lines of investigation are particularly important: whether the licensing condition on deletion is purely semantic and, relatedly, whether mutual entailment is the proper semantic relation. We begin with the first.

2.1 SYNTACTIC ISOMORPHISM

Clearly, invoking both syntactic and semantic conditions on ellipsis is not ideal, especially given how closely semantic composition hews to the syntax. But over the past decade and a half, evidence has steadily accumulated that the possibility and form of sluicing are sensitive to the morphosyntax of the antecedent clause (Merchant 2005, Chung 2005, 2013, Chung et al. 2011). Such observations suggest that sluicing requires a structural (linguistic) antecedent and, concomitantly, that the licensing condition involves information about the antecedent which is not plausibly recoverable from its semantic interpretation. For example, in contrast with VP ellipsis, sluicing does not tolerate voice mismatch (Merchant 2001: 34–35, Chung 2005, Chung et al. 2011, AnderBois 2011b, Chung 2013, Merchant 2013):

[1] Though see Merchant (2004, 2014) for some important qualifications.

[2] Obviously the three alternatives are not mutually exclusive; one can hold that, say, Type 1 and Type 3 ellipses co-exist for a given language or for a given ellipsis-type. See Craenenbroeck (2010), Baltin (2012), Merchant (2014) for recent discussions of this possibility.

(6) a. The candidate was abducted but we don't know who by/by who.
b. *Somebody abducted the candidate, but we don't know who by/by who.
c. Somebody abducted the candidate, but we don't know by who he was abducted.

If actives and passives are semantically equivalent vis à vis licensing ellipsis (as suggested by active-passive mismatches under VP ellipsis (Kehler, 2002:53)), then the sturdy impossibility of (6b) suggests that sluicing requires a licensing condition beyond semantic equivalence, such as one forcing the two clauses to use the same lexical resources composed in the same way (Merchant 2005, 2013, Chung et al. (2011), Chung 2013).

Analogous difficulties arise for cases that have been discussed under the rubric of 'Chung's Generalization'. Chung (2005) observed that bare nominal WH-phrases cannot be sluiced in certain cases in which the antecedent clause lacks a crucial governing preposition. Compare (7), with a prepositional phrase and (8), in which the interrogative phrase is (by inference) the object of a stranded preposition.

(7) a. They're jealous but it's unclear who of.
b. Last night he was very afraid, but he couldn't tell us what of.

(8) a. *They're jealous but it's unclear who.
b. *Last night he was very afraid, but he couldn't tell us what.

Of course, preposition stranding in the absence of ellipsis is unproblematic:

(9) a. They're jealous but it's unclear who [they're jealous of].
b. Last night he was very afraid, but he couldn't tell us what [he was very afraid of].

The puzzle here is why, under a purely semantic licensing condition, (8a-b) cannot be derived from (9a-b). First, in these cases mutual entailment seems to hold, insofar as *They're jealous* can mean *They're jealous of someone*, so ellipsis should be possible here for Merchant (2001). For Type 1 proposals, ellipsis is possible as long as the variable forming the ellipsis site can be anaphorically resolved to a pragmatically salient proposition or issue (a 'question under discussion' (QUD) in the theories of Ginzburg & Sag (2000) and Ginzburg (2012)). But it is likewise reasonable to expect that *They're jealous* would make salient the proposition expressed by, or the issue raised by, a sentence like *They're jealous of someone*. Yet (8) is strongly ill-formed, no matter the discourse context. The contrast between (8) and (7) seems to turn on whether the WH-phrase has access to the particular pleonastic preposition *of* which selects it and is in turn selected by *jealous*. That is, the illformedness of (8) turns on purely formal properties (morphosyntactic) of an antecedent clause.

Thus, while semantico-pragmatic mechanisms are of central importance in sluicing, it seems fair to conclude that the syntactic matching effects documented over the past fifteen years are unexpected given purely semantic or pragmatic understandings of the resolution mechanisms. But, as Merchant (2001:18–25) documents (as evidence for purely semantic licensing, in fact), the morphosyntactic parallelism required is partial and imperfect:

(10) a. I can't play quarterback; I don't even know how.
b. I'll fix the car if you tell me how.
c. I remember meeting him but I don't remember when.

This conundrum—the simultaneous sensitivity of the morphosyntactic condition to lexical content and syntactic structure, alongside its blindness to mismatches in finiteness or lexical category—has given rise to proposals that the isomorphism requirement (whatever its ultimate origin) enforces only skeletal or selective matching (Merchant 2005, Chung 2013).

2.2 SEMANTIC PARALLELISM

It is also unclear whether mutual entailment is the right way to understand the semantic licensing condition. Many cases of sprouting appear to suggest that mutual entailment is not necessary, as illustrated by the examples in (11) (Fox 1999, Chung 2005, Chung et al. 2011, AnderBois 2011b, 2013).

(11) a. She was babbling away, but about what, I have no idea.
b. He finished on time, but with whose help?

But it is at least non-obvious that, e.g., *he finished on time* entails *he finished on time with someone's help*.[3] Though one may postulate implicit correlates (inner antecedents in our terms) within the antecedent for such cases (Merchant 2001, AnderBois 2013) and this may be warranted for *bona fide* implicit arguments (for relevant recent discussion see Bhatt & Pancheva (2006), Landau (2010)), such a move seems more dubious for adjuncts like those in (11) and (10).[4] In addition, this move implies eliminating the distinction between sluicing based on merger and sluicing based on sprouting and that leaves open the question of why there seem to be differences between the two cases in terms of island amelioration (Chung et al. 1995, Yoshida et al. 2013) and in terms of how they are processed (Frazier & Clifton 2000, Dickey & Bunger 2010).

In sum, a substantial body of evidence now exists suggesting that much more syntactic parallelism is required between antecedent and ellipsis site than is expected under purely semantic or pragmatic approaches to the resolution problem. The data seem to argue for the least parsimonious analysis—a Type 3 analysis which incorporates parallel syntactic and semantic licensing conditions. But the status of these two conditions as proposed remains unclear.

3 AN OPPORTUNITY AND A PLAN

We see this as a moment of great opportunity. Research on sluicing (and ellipsis in general) since 1969 has narrowed the space of possible answers to the major questions considerably. At the same time, novel data-sets and tools now open up a new avenue for progress on the fundamental questions: what elided content is, how ellipsis is resolved, and (most fundamentally) why the option of ellipsis exists. To make the most of these opportunities, we need three things:

MISMATCH: To understand the role of syntax in ellipsis resolution, we need a fuller, more systematic map of the dimensions of mismatch between antecedents and ellipsis sites.

SPROUTING: To assess whether there is an implicit IA for sprouted WH-phrases, we need a more systematic understanding of discourses in which sprouting occurs—Was the sluiced content an issue in that context? If not, do such examples bear the hallmarks of accommodation or coercion?

REDUCTION: To understand why ellipsis exists, we must understand the factors which encourage its deployment. Despite their importance, these factors are woefully understudied (we know only of Hardt & Rambow 2001 and Kertz 2013).

Corpus work is a natural vehicle for deepening understanding on all of these fronts; and with current tools, we can extract large-scale data-sets which can be exploited to uncover new patterns, test hypotheses, and train

[3] AnderBois (2011b, 2013) argues that mutual entailment is insufficient, considering cases like **It is not the case that she does not plan to marry anyone, but I don't know who*, which fails, as opposed to *She plans to marry someone, but I don't know who* which succeeds, even though the entailments of the possible antecedents are apparently identical in the two cases.

[4] Fox (1999) treats such cases in terms of accommodation. The challenge here is to limit the role of accommodation so that it does not undo the understanding of all of the cases of illformedness (e.g. (8)) that we have discussed already.

machine learning algorithms. To be useful to theoreticians, such databases require careful, linguistically-sophisticated annotation schemes which target questions of real theoretical interest; when the annotation protocols are well designed, new and unanticipated patterns can emerge which in turn inform new research agendas and lead to the asking of new questions. Clearly, though, the full potential of such a resource can be realized only if the data-base is open and accessible to all (not the property of one researcher or one research-group) and only if the tools with which it can be queried are intuitive and usable by all.

We propose to build such a database for sluicing and to make it accessible to all, in the hope that it can lead to real progress on the kinds of theoretical questions we have laid out here. It is eminently feasible, within the three-year funding window, to construct for sluicing in English a larger and more sophisticated database than currently exists for any ellipsis process in any language. For the particular theoretical questions we focus on in this proposal, for example, it should allow us to truly map the dimensions of possible formal mismatches between antecedent and ellipsis site, to assess the role of implicit content in sluicing, and ultimately, we hope, to provide empirical grounding for the question of when and why speakers deploy sluicing or choose not to.

But of course these questions represent only a small subset of those that might be investigated. One might ask (for instance): what is the role of E-type pronouns in sluicing? Are root sluices fundamentally different from embedded sluices (as one reading of Ginzburg & Sag 2000 might suggest)? Are cross-speaker sluices different in basic ways from same-speaker sluices? What, really, is the truth about island-amelioration under sluicing? What is the range of possible IA's? What is the class of prepositions in English which supports inversion under sluicing (SWIPING as Jason Merchant named it)? And while we are far from suggesting that this methodology will or should replace other well-established methodologies in theoretical syntax and semantics, we ARE suggesting that such a tool could be an invaluable source of new knowledge.

Building such a resources requires (i) expertise in the area of ellipsis, (ii) expertise in the construction of corpora (iii) a pool of sophisticated annotators, and (iv) the funds to support them and to develop the needed infrastructure. The first three we have. The Linguistics Department at UCSC has a long history of engagement with the problems of ellipsis and some of the most important work on the topic has been done by faculty and students associated with the department. As a consequence, there is a deep pool of expertise on the topic to draw on. In particular, a lot of the theoretical work on sluicing has been done at UCSC, by Co-PI McCloskey (a syntactician) and colleagues and graduate students. Co-PI Hardt (who holds a position as Research Associate of the Linguistics Research Center at UCSC in addition to his regular position in Copenhagen) has made contributions of central importance to the semantics and pragmatics of ellipsis and in addition has very substantial expertise in computational linguistics, natural language processing and database development. PI Anand, meanwhile, is in the first place a formal semanticist, with strong research interests and expertise in lexical semantics, in how context guides the process of semantic interpretation, and in the typology of propositional attitude verbs (all crucial for sluicing). Just as important for the present project, however, is the fact that he has expertise and experience (both theoretical and practical) in high-level computational linguistics. Much of his recent research, in fact, involves corpora constructed from detailed annotations of higher-level discourse functions (focus, persuasiveness, and stance). In addition to the core personnel, Sandra Chung, Jorge Hankamer, and William Ladusaw have committed to participating in the project as (un-salaried) consultants, in ways that we will specify in the Research Plan in Section 5 below.

For annotation, we will employ a rich local resource (unique in the world as far as we know): UCSC linguistics undergraduates, who receive an in-depth, hands on training in sophisticated syntactic and semantic analysis, with a particular focus on ellipsis in the more advanced stages, and who are very eager to be involved in large-scale collaborative research projects. Students have found that this kind of sophisticated annotation is at once enjoyable, remunerative, and intellectually engaging. Through involvement in such projects in the recent past, three of Anand's undergraduate annotators have been co-authors in published research, two have

gone on to graduate linguistics programs, and four have begun industry careers in computational linguistics.

We elaborate the plan below; first we place it in context by reviewing similar efforts by other investigators.

3.1 PREVIOUS WORK

As far as we know, there are precisely seven systematic corpus annotations of ellipsis, four focusing on verb phrase ellipsis of various kinds (Hardt 1997, Nielsen 2005, Bos & Spenader 2011, Shahabi & Baptista 2012) and three on sluicing (Fernández et al. 2005, Beecher 2008, Nykiel 2010).

The first large-scale study of verbal ellipsis is due to Hardt (1997), whose central interest is in automatically identifying antecedents in cases of VP ellipsis, based on an examination of 644 instances extracted from the Penn Treebank (Marcus et al. 1993). Nielsen (2005) read through one million words across two corpora and uncovered 1510 instances of VP ellipsis. In addition to the antecedents, he provides intuitive text paraphrases of the resolved content and the general type of mismatch with the antecedent. In a similar effort, Bos & Spenader (2011) manually examined the modals and auxiliaries that license VP ellipsis in the entire Wall Street Journal portion of the Penn Treebank. They find 580 instances which they code for antecedent and a range of other categories. Shahabi & Baptista (2012) are interested in the question: how do different languages cope with the same potential redundancy? They examine the Tehran English Persian Parallel Corpus (Pilevar-Taher et al. 2011), an automatically aligned English-to-Persian parallel corpus), find 10,515 instances of VP ellipsis in English and determine how many of those instances are not elided in Persian.

For sluicing, there are three principal efforts, all with very particular goals. Nykiel (2010) traces the relative frequency of sprouting and merger from Old English to Present Day English. Beecher (2008) extracts 3,000 instances of swiping in an effort to understand which prepositions allow it. Fernández et al. (2005) focus on root sluices, extracting 5343 from the BNC (BNC Consortium 2007), and annotate 10% of those for pragmatic function. While all of these efforts are extremely valuable, none aim to provide the kind of corpus (large, systematic, theoretically-guided, and—crucially—openly accessible) that is our goal.

3.2 OUR PLAN

Our own efforts began in the summer of 2013, when a group of faculty, graduate students and undergraduates initiated the project. We used TGREP2 (Rohde 2001–5), a treebank search utility, to extract complement sluices from an automatically parsed version of the *New York Times* subset of the English Gigaword Second Edition corpus (Graff et al., 2005). After some culling of irrelevant and mis-parsed cases, this yielded 4100 genuine examples and this set formed the basis for our initial annotation effort. English Gigaword is large (and therefore more likely to contain interesting phenomena) and is widely in the NLP community, playing well with conventional parsing technology and serving as the basis for other annotation resources we could potentially leverage (and vice versa).

During the summer months and into the Fall, we set about (i) developing a set of annotation protocols (ii) using those protocols to build an initial database using a version of the BRAT annotation tool (Stenetorp et al. 2012) which had been heavily modified by PI Anand. Apart from the PIs, the initial team consisted of six additional annotators—one doctoral student and five undergraduate majors who had just completed a regularly-offered upper-division course on ellipsis. Initial progress was rapid. By October 2013, we had annotated 417 sluices in seven rounds of annotation (all with biased sampling to find potentially problematic cases) with good levels of inter-annotator agreement (see below). We invite and encourage reviewers to view these initial results here, bearing in mind that what they are seeing here is the annotator view, NOT the planned user interface. Our work followed agile annotation (Alex et al. 2010), where code-book development and

annotation proceed simultaneously, and the work was therefore done interactively and collaboratively and decisions about how to treat difficult cases were adjudicated in group discussions. The resulting guidelines were codified into an initial annotation manual in the Winter and Spring of 2014.

A crucial property of the data-set is that each example comes with a substantial context window (preceding and following). We found that a radius of five sentences was needed for reliable annotation; even when the antecedent was nearby (typically in the preceding sentence), determining the proper antecedent scope and ellipsis resolution often involves understanding the discourse dynamics—in particular the questions under discussion at the point sluicing is deployed. That information is retrievable for each example. This is crucially important, since ellipses are, of course, extraordinarily context-dependent, both in terms of their relative wellformedness and in terms of their interpretation. One of the ways in which corpus work can help us make progress is exactly in providing the relevant context for each crucial example. It is not so difficult for a trained investigator in syntax or semantics to invent crucial examples in isolation; what is often required in work on ellipsis, however, is not just the example but the kind of contexts (often large and complex) in which it might be used. It is not easy to conjure up such rich contexts, but they are critically important.

3.3 ANNOTATION SCHEME

Crucial to the success of the project will be the annotation scheme. In drawing up protocols, we have drawn heavily on the existing theoretical literature on sluicing. Our choices have been informed by the conviction that an annotation scheme which tries to avoid all theoretical commitment will be of little use to future investigators. At the same time, however, the scheme must be sufficiently catholic that it will be useful to researchers of very different theoretical and analytical persuasions. It must also strike a balance between sophistication and usability for end-users and it must be implementable by our annotators.

Our current annotation scheme code-book was shaped by these design goals; it can be viewed here. Each example is annotated with four obligatory tags: the ANTECEDENT, the SLUICE—including a plain-text paraphrase of the elided content—the main PREDICATE of the Antecedent, and the INNER ANTECEDENT, if there is one. Several tags are also tagged with certain important taxonomic features (type and extension of IA, type of implicit argument sprouted, and morphosyntactic mismatch). In addition, each example may bear six optional tags. Two correspond to cases where there are several possible antecedents. In the case of ALTERNATIVE ANTECEDENT, we observed several cases of antecedent "sandwiching", in which the sluice is buttressed by roughly synonymous potential antecedents, as in (12). ELLIPSIS ANTECEDENT is used in cases where the antecedent for a sluice is itself elliptical (so far, exclusively via VP ellipsis; see (13)).

(12) We lost our focus a little bit somewhere. I don't know where. But we lost it. [27861]

(13) a. 'It's difficult to get the black church to deal with the issues, ' said Campbell after Elders's speech. 'It's not that we don't want to, it's that we don't know how. [21666]

b. WOLF: Have you changed over time as performers? BATES: You're bound to have, but you don't always know how. [54048]

Two additional tags deal with interpretive differences between antecedent and ellipsis site. E-TYPE marks indefinite material in the antecedent that is interpreted anaphorically in the ellipsis site, as in (14).

(14) She said that she would issue a written ruling as soon as possible, but did not say when. [35291]

In contrast, IGNORE is used to mark material in the antecedent that does not seem to be part of the interpretation of the ellipsis site (parenthetical material, for example, or additive particles, or high-attaching adverbs).

MANDATORY TAGS

SLUICE: sluice site.

- ISLAND: whether sluice 'crosses' an island
- Mismatches [Finiteness, Tense, Person, Case, Subject Overtness, Additional Words, Other]

ANTECEDENT: intuitive fill for Ellipsis Site

PREDICATE: main predicate for clause in A.

- SPROUTED PARAMETER TYPE: [Adjunct, Passive Subject, Other]

INNER ANTECEDENT: material in A replaced or elaborated on by WH-phrase.

- TYPE [Indefinite, Definite, Pronoun, Strong Quantifier, WH-phrase, Name, Disjunction, Temporal/Locative, Degree/Extent]
- EXTENDED: how WH-phrase extends IA [Possessor, Focus Particle, Type, Implicit Object, Temporal, Locative, Other]

OPTIONAL TAGS

ELLIPSIS ANTECEDENT: A is elided

ALTERNATIVE ANTECEDENT: Secondary A

E-TYPE: Indefinite in A that is anaphoric in ES

DOMAIN: Domain restrictions for IA

DEGREE MODIFIER: Modifier of a Degree Sluice

IGNORE: Material not retained in ES

Figure 1: Abridged Sluicing 1.0 Tagset

The final two tags, DOMAIN and DEGREE MODIFIER, are used to describe restrictions on the WH-remnant provided by material surrounding the Inner Antecedent. The issues raised here are interesting, and they are discussed in section 4. A complete overview of the current tag-set is given in Figure 1 on the next page.

3.4 BUILDING ON THE FOUNDATION

This is the foundation we intend to build on to construct the resource we described earlier (see the detailed plan in section 5). The annotation scheme will be under constant revision as the work proceeds. In its present form it does certain things well (see section 4 for some evidence) but it should do more. Two particular opportunities suggest themselves. First, at present, there is hardly any pragmatic annotation, although this is needed (see section 5 for some ideas). In addition, the representational logic used for resolving the ellipsis is of central importance. We found free-text paraphrase helpful for identifying mismatches, but we would like the final scheme to move beyond bare text to a system which marks (dis)correspondences between antecedent and sluice. Existing annotation tools make this difficult, since those that allow insertion of new markables (e.g., MMAX2 (Müller & Strube, 2006)) completely alter the document, making inter-annotator comparison difficult. We will thus need to build a new annotation tool which handles this more gracefully. We also need to better handle ambiguity and also the theoretically interesting case of complex coordinated sluices like (15) or the equally interesting split antecedent cases like (16). Such features will be crucial for annotating implicit content in general, and we expect that such a tool, designed for theoretical linguists, will be of general value.

(15) To those who have faulted him for not lobbying aggressively for permanent trade relations for China, he said he had called "a bunch" of members of Congress, but would not say how many or whom. [89868]

(16) Bill Gates, in the Wall Street Journal, is calling it his "favorite business book." (He says it's Warren Buffett's favorite business book, too.) It's easy to see why. *The New Yorker*

While building this resource, we will exploit it in three ways. First, we will pursue the the empirical investigations discussed in section 2; second, we plan to address the fundamental question of when and why ellipsis is used. To do this, we propose to build a parallel database of contexts in which sluicing *could* have been used but was not (see section 5). From comparison of such cases with actual instances of sluicing, we should be able to make inferences about the factors which favor or disfavor reduction at the ellipsis-site. Finally, looking farther ahead (see section 6, we will explore the utility of this large-scale data-set as a basis for machine learning for sluicing, as well as the relative viability of crowd-sourcing some portion of our annotations.

4 Grounds for Confidence

In our work so far we have been surprised and encouraged by (i) the level of sophistication we were able to achieve in the annotation (ii) the skill and dedication of the undergraduate, and (iii) how rewarding they found the work to be. By round 5, they reported being able to complete 15-20 annotations per hour and we had achieved good levels of inter-annotator agreement. The interannotator agreement for the tags across the rounds is provided in Table 1. As the tags all mark text spans, we use Krippendorff's continuum metric (Krippendorff, 1995) (a special case of Krippendorff's α (Krippendorff, 2014) for spans).

Most of the agreement gains come from conventions about boundaries (e.g., when ignored material at clause-edge should be marked Ignore vs. excluded from the Antecedent, what the predicates of copula and existential sentences are), but some involved actual instruction of the undergraduates (e.g., correct cases of EType), and others involved implicit learning (e.g., what counted as the "real" IA in an expression). Most of the optional tags arose from disagreements that, in discussion, revealed fundamental oversights or confusions in our original taxonomization of sluicing cases, and the revisions they led to yielded corresponding increases in agreement for Antecedent and IA.

	Round				
Tag	2	3	4	5	6
Ante	.83	.67	.73	.78	.88
Pred	.92	.56	.85	.85	.85
IAnte	.72	.58	.60	.74	.78
Elided				.94	.94
AltAnte				.66	.78
EType	.21	.32	.67	.80	.87
Dom					.66
DegMod				.56	.72
Ignore			.43	.74	.78

Table 1: Pilot Inter-Annotator Agreement.

Even though our current set of annotated examples is tiny by comparison with our ultimate goal, we are encouraged by the fact we have already encountered phenomena of real theoretical interest.

In particular, several kinds of mismatch between antecedent and ellipsis site have turned up which have gone undiscussed or underdiscussed in previous work. Here we offer some examples, as an illustration of the potential for discovery that we think our resource holds out.

4.1 Modal mismatches

Since Merchant (2001), it has been known that a finite clause can antecede a nonfinite sluice, triggering attendant realis differences, as in (10a) above. But we have also found many (40) examples of the reverse pattern, where a non-finite (or modal) antecedes a sluice. In 30 of these cases, the precise modality intended inside the sluice is difficult to pin down. In (17), for example, is the intended modal here a simple future, or a

future-oriented modal (if so, of what flavor?)? In such cases, the overt intensional operator of the antecedent is paralleled by an underspecified modal in the interpretation of the ellipsis (which we gloss here and in our annotations as MODAL). (Here, and below we use angled brackets to give an informal indication of the interpretation of elided material.)

(17) "I want to return (to Peru) some day , but I don't know when < I MODAL return to Peru> … " [117524]

(18) "Basically ," Atwater said , "they asked me in two or three games for four or five series to sit down on third-down plays, and I didn't understand why < I MODAL sit down on third-down plays>." [40784]

(19) Texas A&M coach Tony Barone unabashedly predicted that despite some key player losses from the team that led the Southwest Conference race most of last season, the Aggies could be better than a year ago. He just forgot to say when <the Aggies MODAL be better than a year ago>.[88489]

In advance of further analysis, we hesitate to offer generalizations about what constraints govern the interpretation of this underspecified modal. But we note that this is a case in which there is no expression of English which accurately renders the interpretation of the elided material. We further speculate that the existence of this underspecified modal element will have interesting implications for our understanding of the structure of modals in English, and also, of course, for our understanding of parallelism conditions on ellipsis.

4.2 COMPOUND INNER ANTECEDENTS

Several of our novel phenomena emerged originally as cases of annotator confusion, including the following:

(20) Despite my inclination toward procrastination, I am determined to send holiday cards this year. It doesn't much matter which holiday. [106579]

This example emerged as a problem during annotation precisely because it is unclear what the shape of the analysis is—what the Ellipsis Site is, what the Antecedent is, how they correspond—and yet all annotators agreed it is grammatical. Three analyses of the ellipsis site are possible: that the WH-phrase is sprouted off *holiday cards*; that it is extracted from the compound nominal *holiday cards*, violating numerous constraints on extraction; or that it is extracted from an elided cleft 'pseudo-sluice' (as in (21c)).

(21) a. It doesn't much matter which holiday <I send holiday cards for>
b. It doesn't much matter which holiday <I send [__ cards]>
c. It doesn't much matter which holiday <it is that I send holiday cards for>

Of these options, both the sprouting and compound nominal cases are empirically novel. If sprouting, it would violate Chung's generalization, and should be as ill-formed as those in (8) (e.g., **They're jealous but it's unclear who.* But this seems wrong, since all annotators felt that without *holiday*, the sluice was substantially degraded. (Note that the same argument can be leveled against the pseudo-sluice analysis.) Assuming that the modifier is (in some sense) indefinite, the compound analysis predicts this contrast. What complicates such a picture is that this indefinite must escape the scope of the intensional quantification in the compound, and this is a matter of pragmatics, as the constrast in (22) shows:

(22) a. He's missing a piano bench, but he didn't tell me (for) which piano.
b. He just finished making a piano bench, but he didn't tell me *(for) which piano.

If the compound analysis is correct, there are issues for the analysis both of compounds and of IA's.

4.3 Degree Expressions

Among our most vexing (and interesting) cases for annotation were degree sluices, underdiscussed in the theoretical literature, but very common in our data (25 cases). A degree WH-phrase (like *how much*) may have no overt IA, as in (23), or may have as IA a vague indefinite extent, as in (24).

(23) a. They said this would save the government money, though they could not yet say how much <this would save the government money>. [2753]

b. The review, Gilligan acknowledged, delayed the issuance of the notice about Strandflex, but she said she could not estimate by how much <the review delayed the issuance of the notice about Strandflex>. [60122]

(24) a. The Atlanta-based company said Thursday that operating profit would be "substantially below" analysts' estimates but didn't specify how much <operating profit would be below analysts' estimates>. [104088]

b. But Thursday the market for other California municipal bonds recovered a bit. "It's difficult to say how much <the market for other Californa municipal bonds recovered>, because … " [35463]

For our annotators, the question was: what is the IA in cases like (24)? The apparent answer is that the IA's are the vague indefinite extents *substantially* and *a bit*. But these elements are optional and in their absence sluicing with *how much* remains possible, much as in (23b). And in such cases, the implicit indefinite degree quantifier still contributes a restriction on the domain of the degree WH-expression of the sluice in just the same way that overt IA's routinely do. But that in turn suggests that the 'real' IA's for such cases are not *substantially* or *a bit*, but rather implicit degree expressions which are modified by *substantially* or *a bit*. That decision in turn suggests a similar analysis for cases like (23), which are similar in all relevant respects.

There is a practical question of annotation here. But as is often the case, annotation dilemmas highlight theoretical puzzles. For the annotation question, our approach (after many false starts) has been to assume that the IA for (24b) is an implicit degree parameter of the verb *recover*, and that *a bit* serves as a Degree Modifier of that parameter. We made this choice for annotational simplicity, but it is very clear that there are important questions of theory and analysis at stake here, with implications at least for the distinction between merger and sprouting. Cases like those in (23) would naturally be taken to be sprouting cases, while those in (24), because there is an overt indefinite, would naturally be taken to be cases of merger. But that bifurcation obscures important (semantic) commonalities between the two kinds of cases, and suggests once more how useful sluicing can be as a probe for implicit content. And since such cases suggest that at least some apparent cases of sprouting need to be analyzed in terms of implicit IA's, they force the question again of whether or not such interpretations are generally correct—a position which would in turn have important ramifications for theories of implicit content more generally. Vexation for annotators often signals phenomena of particular theoretical interest.

5 Research Plan

We aim in the award period to construct a corpus of roughly 27,500 instances of sluicing, across several genres. The resulting corpus will contain the following sections (in order of creation): (i) 10,000 sluices from large-scale monologic text corpora, (ii) 10,000 sluices from dialogic corpora, (iii) the sluice-like (pre-

sluices and clefted question) elements from the text in (i), and (iv) 2,500 fragment answers from the dialogic text in (ii).

In the first year, our aim is to exhaustively annotate the sluices in large-scale written corpora. This includes the 4,100 sluices already identified, as well those in the rest of Gigaword and other large-scale corpora, (e.g., the ANC (Ide & Suderman, 2004) and GloWBE (Davies, 2013)). We will also examine smaller reference corpora (e.g., The Penn Treebank and MASC (Passonneau et al., 2012)), which already have additional rich semantico-pragmatic annotations (e.g., discourse relations, coreference, argument structure) we will use to add pragmatic factors to our annotation scheme. We believe our 10,000 sluice aim is conservative: We will training five undergraduate annotators. Assuming that each example is annotated by two annotators, and that each annotator annotates 14 sluices/hour and for 296 hours/yr yields 10,360 annotations.

We will then extend in two directions. First, we will turn to spoken language, drawing from the conversation sections of the ANC and BNC, as well as scripts and closed captions/subtitles for movies and television shows (via the Internet Movie Script Database (http://www.imsdb.com/) and OpenSubtitles (http://www.opensubtitles.org/en/search)). Based on Fernández et al. (2005), we expect to find a larger number of root sluices in this genre than the newswire that predominates the data from the first year.

We will also find and annotate two sluice-like constructions: a) 'pre-sluices', counterparts of sluices where, for Merchant (2001), clausal material is unelided and b) clefted questions, the overt counterparts of pseudo-sluices. To uncover these, we will need to find all WH-questions in the text, and then filter those to one with nearby 'antecedents.' These tasks will initially be done by machine, and hence this work will piggyback on our attempts to automatically find the antecedents of sluices. The results will then feed a bootstrap procedure: (i) annotators correct system, (ii) system is updated, (iii) system generates new candidates, etc.

In the final year, we will expand the scope again. First, in analogy with the effort of Shahabi & Baptista (2012), we will consider how sluice and sluice-like constructions are realized cross-linguistically by examining their translations in parallel text corpora. We will likely use the Europarl corpus (Koehn, 2005), a 300 million word corpus of the proceedings of the European Parliament, translated into 11 languages; we will concentrate on English and perhaps also on Spanish depending on the availability of local expertise. Second, we will extend our investigation of root sluices to cover other types of fragments, including fragment answers, which have played a central role in the debate about whether sluicing involves any syntactic component (Stainton, 2006, Ginzburg & Sag, 2000).

For each new corpus and genre, we will need to both gather the relevant data and perform initial annotation to modify or extend the coding scheme. This will be one of the principal aims of the graduate students working on the project. A skeletal schedule of how data gathering, annotation design, and annotation will proceed is provided in the first three columns of Figure 5.

Alongside the principal annotation task, we will explore the viability of matching our annotations by machine and by relative novices. Machine learning is important in two respects. First, it will allow more efficient identification of our elliptical phenomena; the resulting algorithms should help theorists find sluice and sluice-like material in additional data. Second, the resolution to elliptical content is currently difficult for machines (see section 6). Demonstrating the utility of this resource for that task will help argue for extending this enterprise beyond sluicing and beyond English. We plan to pursue this gradually, first focusing on identifying cases of sluicing, then finding antecedents and inner antecedents, and then tackling resolution. Figure 2 provides a schedule of how we will proceed.

Beginning the second year, we will also undertake a series of crowdsourcing experiments to test whether the various subtasks our undergraduate annotators perform can be done by relative non-experts. There is an astonishingly large amount of work on annotating linguistic data via crowdsourced novices, but it is still unclear whether annotation tasks requiring detailed linguistic knowledge (like syntactic or semantic structure)

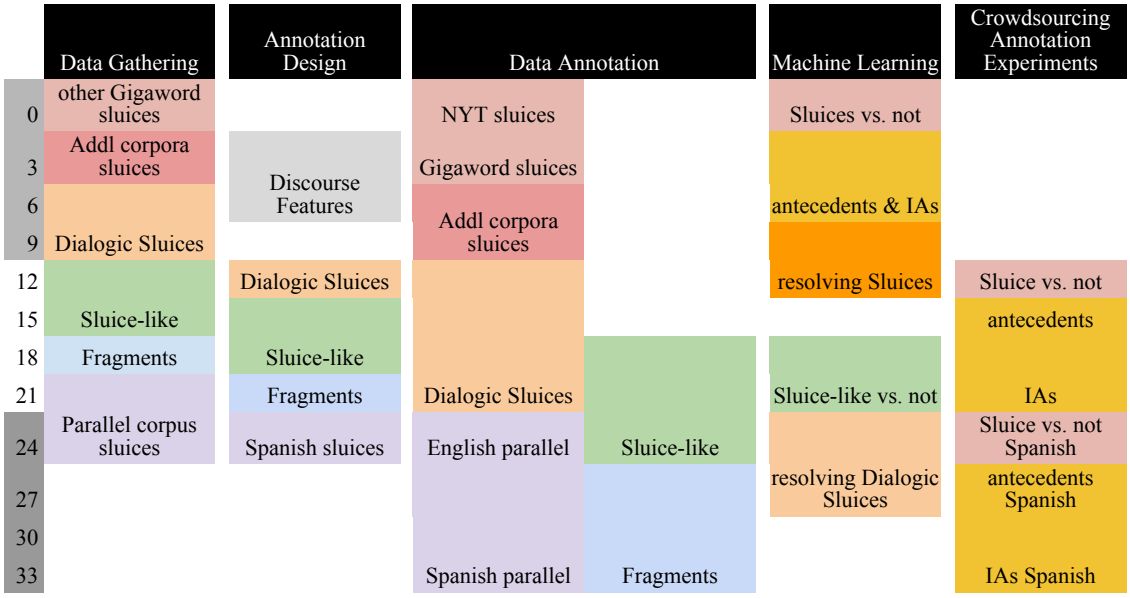

Figure 2: Timeline of work, in three month intervals.

can be done by non-experts. This question is important because training expert annotators is often seen as expensive and inconvenient. Beyond novices, our plan is to determine to what extent we can iteratively *train* crowdsourced workers to perform some of the harder tasks (see Figure 5 for the timeline). This is not often done, but we suspect that it is possible. In Y2, we will attempt to train five annotators to the level of our undergraduates. If successful, in Y3 we will aim for 25 annotators, across English and Spanish.

Our plan includes milestones for reflection and data release. Three times each year, Co-PI Hardt will spend one week at UCSC. These visits will provide an opportunity for the team as a whole to present progress reports to other team members and to solicit feedback from our consultants (Chung, Hankamer, and Ladusaw). In addition, we will release our results (annotations, codebooks, and computational models) yearly. Making this resource accessible to linguists is a central aim of this project. Hence, as part of the release process, we will survey researchers working on ellipsis and conduct a usability study (with five Bay Area graduate students unrelated to the project), on the corpus's search interface to the corpus.

6 Future Prospects

By the end of the award period, a tangible product should have emerged from our work—the annotated corpus we have described here. We are confident that that resource will be of use to many different kinds of researchers, but its limitations are also apparent. It will be devoted to a single ellipsis-type and its crosslinguistic coverage will at that point still be very limited. However, we will also emerge with the infrastructure and best practices for annotating implicit content more generally. In particular, a natural extension of the project is to include other ellipsis types and subsentential fragments and, crucially, to go much farther beyond English in crosslinguistic coverage than is possible at present.

Our focus in this proposal has been on linguistic theory, but we believe that there will also be substantial interest in the NLP community in the resources we aim to build. In a recent interview, Ronald Kaplan described the new centrality of and current challenges for language-based technologies, especially for mobile

devices. With the rapid progress in speech recognition and synthesis, the hard problems are now at higher levels, in 'that space between the way that ordinary people say things and the complex functional interfaces that engineers are building into all these devices'. Dialog modeling at Google, for example, is focusing on the contextual representation necessary to enable conversational search, following up on the results of an initial query with a second. Such follow-up queries will normally be full of anaphora, ellipses, and subsentential fragments of all kinds (including sluices) because ordinary conversations are full of such elements. With these realities in mind, our long-term plan is to build on the work of the current project and use it as a basis for a larger application (perhaps collaborative) at a later point to the CISE Computing Research Infrastructure (CRI) initiative of the CISE directorate of the National Science Foundation, a program whose express goal it is to foster the creation of new kinds of research infrastructure. We contend that the resources that could become available under such a program have the potential to bring large benefits both to the language sciences community and to the language engineering community.

7 Broader Impacts

We see our project as having two kinds of broader impacts—educational and technological.

Our plan is for undergraduate students at UCSC to do much of the annotation work. The five undergraduate annotators in our pilot have found the work to be engaging and rewarding, and they are eager to continue. Doing annotation requires that they acquire a sophisticated knowledge of the phenomenon, that they become familiar with computational tools (many of them quite advanced), and that they work collaboratively and efficiently within a group. Such experience prepares them well for futures in academia or the technology sector. Our experience with former students now in that sector suggests that annotation is now one of the principal tasks that those with training in linguistics are being sought for. Beyond all that, students spoke of the excitement and joy that they felt in participating in a central way in the process of scientific discovery.

UCSC, which in 2015 will become a Hispanic-Serving Institution, serves a large and very diverse population of students, many from communities whose first language is not English. When we extend the reach of our project beyond English, we will be relying on the language-expertise of these students and providing those who work on the project with a valuable educational experience, as well as a form of linguistic empowerment. The onus is on us, of course, to ensure that participating students are trained and mentored. Most of this will be done within the framework of the project (in group meetings and training sessions), but we also anticipate the work synergizing with our existing curriculum in syntax and semantics and with courses in computational, experimental, and statistical methods, where access to the corpus would be of great benefit.

In addition to these educational benefits, we believe that the project has the potential to speed certain important technological innovations of the kind described in section 6. Making this possible means computationally solving the resolution problem, a problem at present much better understood by linguists than by engineers. We believe that the provision of richly annotated corpora of the kind we envisage here has the potential to narrow this important technological gap.

8 Conclusion

Whether or not we succeed on the most ambitious questions, we are confident that with the help of NSF funding we can make available an important new resource and make real progress on the important subsidiary questions (the typology of mismatches, the typology of sprouting and so on). More importantly, although our discussion has focused on fairy particular questions, our project should serve as a useful guide for the building of large-scale, richly annotated databases for a wide range of important theoretical issues.

References

Alex, Bea, Claire Grover, Rongzhou Shen & Mijail Kabadjov. 2010. Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs. In *Proceedings of the fourth linguistic annotation workshop* LAW IV '10, 29–37. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1868720.1868724.

AnderBois, Scott. 2011a. *Issues and alternatives*: University of California Santa Cruz dissertation.

AnderBois, Scott. 2011b. Sluicing as anaphora to issues. In Dan Li & David Lutz (eds.), *SALT 20, proceedings from Semantics and Linguistic Theory 20*, Available at: http://elanguage.net/journals/salt/issue/view/220.

AnderBois, Scott. 2013. The semantics of sluicing: Beyond truth conditions. Manuscript, Brown University. To appear in *Language*.

Baker, Collin. 2008. FrameNet, present and future. In Jonathan Webster, Nancy Ide & Alex Chengyu Fang (eds.), *The first international conference on global interoperability for language resources*, City University Hong Kong: City University.

Baker, Collin, Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the conference*, 86–90. Montreal, Canada.

Baltin, Mark. 2012. Deletion vs. proforms: an overly simple dichotomy? *Natural Language and Linguistic Theory* 30. 381–423.

Barker, Chris. 2013. Scopability and sluicing. *Linguistics and Philosophy* 36. 187–223.

Beecher, Henry. 2008. Pramatic inference in the interpretation of sluiced Prepositional Phrases. In *San diego linguistic papers*, vol. 3, 2–10. La Jolla, California: Department of Linguistics, UCSD.

Bhatt, Rajesh & Roumyana Pancheva. 2006. Implicit arguments. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, vol. 2, 558–588. Oxford: Blackwell.

Biber, D., E. Finegan & D. Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie & P. Schneider (eds.), *Creating and using English language corpora*, 1–13. Amsterdam: Rodopi.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Bos, Johan & Jennifer Spenader. 2011. An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation* 45. 463–494.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó & Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th international conference on language resources and evaluation (LREC 2006)*, Genoa, Italy.

Büring, Daniel. 2003. On d-trees, beans, and accents. *Linguistics and Philosophy* 26. 511–545.

Chomsky, Noam. 1972. Some empirical issues in the theory of Transformational Grammar. In Stanley Peters (ed.), *Goals of linguistic theory*, 63–130. Englewood Cliffs, New Jersey: Prentice Hall.

Chomsky, Noam. 2001. Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*, 1–52. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 2008. On phases. In Robert Freidin, Carlos Otero & Maria Luisa Zubizaretta (eds.), *Foundational issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud*, 133–166. Cambridge, Mass.: MIT Press.

Chung, Sandra. 2005. Sluicing and the lexicon: The point of no return. In Rebecca Cover & Yuni Kim (eds.), *BLS 31, Proceedings of the Thirty-First Annual Meeting of the Berkeley Linguistics Society*, 73–91. Berkeley, Calif.: 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. 2004. *Restriction and saturation*. Cambridge, Mass.: MIT Press.

Chung, Sandra, William Ladusaw & James McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics* 3. 239–282.

Chung, Sandra, William Ladusaw & James McCloskey. 2011. Sluicing(:) between structure and inference. In Rodrigo Gutiérrez-Bravo, Line Mikkelsen & Eric Potsdam (eds.), *Representing language: Essays in honor of Judith Aissen*, 31–50. Santa Cruz, California: California Digital Library eScholarship Repository. Linguistic Research Center, University of California Santa Cruz.

Craenenbroeck, Jereon van. 2010. *The syntax of ellipsis: Evidence from Dutch dialects*. Oxford: Oxford University Press.

Culicover, Peter & Ray Jackendoff. 2005. *Simpler syntax*. Oxford and New York: Oxford University Press.

Darymple, Mary, Stuart M. Schieber & Fernanda C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy* 14. 399–452.

Davies, Mark. 2008. The Corpus of Contemporary American English: 450 million words, 1990–present. Available online at http://corpus.byu.edu/coca/.

Davies, Mark. 2013. Corpus of global web-based English: 1.9 billion words from speakers in 20 countries. Available online at http://corpus2.byu.edu/glowbe/.

Dickey, M. W. & A. Bunger. 2010. Comprehension of elided structure: Evidence from sluicing. *Language and Cognitive Process* 26. 63–78.

Farkas, Donka & Kim Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics* 27. 81–118.

Fernández, Raquel, Jonathan Ginzburg & Shalom Lappin. 2005. Automatic bare sluice disambiguation in dialogue. In *Proceedings of the IWCS-6 (Sixth International Workshop on Computational Semantics)*, 115–127. Tilburg, the Netherlands. Available at:
http://www.dcs.kcl.ac.uk/staff/lappin/recent_papers_index.html.

Fiengo, Robert & Robert May. 1994. *Indices and identity*. Cambridge, Mass.: MIT Press.

Fox, Danny. 1999. Focus, parallelism and accommodation. In Tanya Mathews & Devon Strolovich (eds.), *SALT IX, proceedings from Semantics and Linguistic Theory IX*, Department of Linguistics, Cornell University: CLC Publications.

Fox, Danny & Howard Lasnik. 2003. Successive-cyclic movement and island repair: The difference between sluicing and VP-ellipsis. *Linguistic Inquiry* 34. 143–154.

Frazier, Lyn & Charles Clifton. 2000. Parsing coordinates and ellipsis. *Syntax* 4. 1–22.

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In J. Seligman (ed.), *Language, logic, and computation*, Stanford, Calif.: CSLI Publications.

Ginzburg, Jonathan. 2012. *The interactive stance: Meaning for communication*. Oxford: Oxford University Press.

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

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

Grebenyova, Lydia. 2006. Sluicing puzzles in Russian. In *Proceedings of the Annual Workshop on Formal Approaches to Slavic linguistics 14 (FASL 14)*, .

Grinder, John & Paul M. Postal. 1971. Missing antecedents. *Linguistic Inquiry* 2. 269–312.

Hacquard, Valentine & Alexis Wellwood. 2012. Embedding epistemic modals in English: a corpus-based study. *Semantics and Pragmatics* 5.4. 1–29. http://dx.doi.org/10.3765/sp.5.4.

Hankamer, Jorge. 1979. *Deletion in coordinate structures*. New York: Garland Publishing.

Hankamer, Jorge & Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry* 7. 391–428.

Hardt, Daniel. 1993. *Verb phrase ellipsis: Form, meaning and processing*: University of Pennsylvania dissertation.

Hardt, Daniel. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 32. 525–541.

Hardt, Daniel. 1999. Dynamic interpretation of Verb Phrase ellipsis. *Linguistics and Philosophy* 22. 187–221.

Hardt, Daniel & Owen Rambow. 2001. Generation of VP ellipsis: A corpus-based approach. In *ACL 01: Proceedings of the 39th annual meeting on association for computational linguistics*, 290–297. Assocation for Computational Linguistics. doi:10.3115/1073012.1073050. http://dl.acm.org/citation.cfm?id=1073012&picked=prox.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*: University of Massachussets, Amherst dissertation.

Heim, Irene. 1997. Predicates or formulas? evidence from ellipsis. In Aaron Lawson (ed.), *SALT VII, proceedings from Semantics and Linguistic Theory VII*, 197–221. Ithaca, New York: Cornell University, CLC Publications.

Ide, N. & K. Suderman. 2004. The American National Corpus first release. In *Proceedings of the fourth language resources and evaluation conference (LREC)*, 1681–84. Lisbon, Portugal. Project website: http://www.anc.org/about/.

Kehler, Andrew. 2002. *Coherence in discourse*. Stanford, Calif.: CSLI Publications.

Kertz, Laura. 2013. Discourse expectations and the grammar of ellipsis. Manuscript, Department of Cognitive, Linguistic and Psychological Sciences, Brown University.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation . In *MT Summit X: Proceedings of the Tenth machine translation summit*, Phuket Island, Thailand, http://www.mt-archive.info/MTS-2005-TOC.htm. Project website: http://www.statmt.org/europarl/.

Krippendorff, Klaus. 1995. On the reliability of unitizing continuous data. In P. V. Marsden (ed.), *Sociological methodology 1995*, vol. 25, Cambridge, Massachusetts: Blackwell.

Krippendorff, Klaus. 2014. *Content analysis: An introduction to its methodology*. Thousand Oaks, California: Sage Publications 3rd edn.

Landau, Idan. 2010. The explicit syntax of implicit arguments. *Linguistic Inquiry* 41. 357–388.

Lappin, Shalom. 1999. An HPSG account of antecedent-contained ellipsis. In Shalom Lappin & Elabbas Benmamoun (eds.), *Fragments: Studies in ellipsis and gapping*, 68–97. Oxford and New York: Oxford University Press.

Lasnik, Howard. 1999. On feature strength: Three minimalist approaches to overt movement. *Linguistic Inquiry* 30. 197–217.

Lasnik, Howard. 2001. When can you save a structure by destroying it? In Minjoo Kim & Uri Strauss (eds.), *Proceedings of the North East Linguistic Society 31*, 301–320. Amherst, Mass.: GLSA.

Lasnik, Howard. 2009. Island repair, non-repair, and the organization of the grammar. In Kleanthes K. Grohmann (ed.), *InterPhases*, 339–353. Oxford and New York: Oxford University Press.

Levin, Beth. 2003. Objecthood and object alternations. Handout from a talk given at the Department of Linguistics, UCLA, May 2nd 2003. Available at: http://www.stanford.edu/~belevin.

Levin, Lori. 1982. Sluicing: A lexical interpretation procedure. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 590–654. Cambridge, Mass.: MIT Press.

Manetta, Emily. 2005. *Wh*-expletives in Hindi-Urdu: The vP phase. Presented at the January 2006 meeting of the Linguistics Society of America, Albaquerque, New Mexico.

Manetta, Emily. 2006. *Peripheries in Kashmiri and Hindi-Urdu*: University of California, Santa Cruz dissertation.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19. 313–330.

Merchant, Jason. 1999. *The syntax of silence: Sluicing, islands, and identity in ellipsis*: University of California Santa Cruz dissertation.

Merchant, Jason. 2001. *The syntax of silence: sluicing, islands, and the theory of ellipsis*. Oxford and New York: Oxford University Press.

Merchant, Jason. 2002. Swiping in Germanic. In C. Jan-Wouter Zwart & Werner Abraham (eds.), *Studies in comparative Germanic syntax*, 289–315. Amsterdam: John Benjamins.

Merchant, Jason. 2004. Fragments and ellipsis. *Linguistics and Philosophy* 27. 661–738.

Merchant, Jason. 2005. Revisiting syntactic identity conditions. Presented at the Workshop on Identity in Ellipsis, UC Berkeley, October 8, 2005.

Merchant, Jason. 2006a. Sluicing. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, 269–289. Oxford: Blackwell.

Merchant, Jason. 2006b. Small structures. In Ljiljana Progovac, Kate Paesani, Eugenia Casielles & Ellen Barton (eds.), *The syntax of nonsententials: Multidisciplinary perspectives* Linguistics Today Series, 73–91. Philadelphia and Amsterdam: John Benjamins.

Merchant, Jason. 2008. An asymmetry in voice mismatches in VP-ellipsis and pseudogapping. *Linguistic Inquiry* 39. 169–179.

Merchant, Jason. 2009. Phrasal and clausal comparatives in Greek and the abstractness of syntax. *Journal of Greek Linguistics* 9. 134–164.

Merchant, Jason. 2013. Voice and ellipsis. *Linguistic Inquiry* 44. 77–108.

Merchant, Jason. 2014. Gender mismatches under nominal ellipsis. *Lingua* To appear.

Merchant, Jason. forthcoming. Variable island repair under ellipsis. In Kyle Johnson (ed.), *Topics in ellipsis*, Cambridge: Cambridge University Press.

Müller, Christoph & Michael Strube. 2006. Multi-level annotation of linguistic data with mmax2. In *Corpus technology and language pedagogy: New resources, new tools, new methods*, .

Nielsen, Leif. 2005. *A corpus-based study of verb phrase ellipsis identification and resolution*: King's College London dissertation.

Nykiel, Johanna. 2010. Whatever happened to Old English sluicing. In Robert A. Cloutier, Anne Marie Hamilton-Brehm & Jr. William A. Kretzschmar (eds.), *Studies in the history of the English language V: Variation and change in English grammar and lexicon: Contemporary approaches*, 37–59. Walter de Gruyter.

Passonneau, Rebecca, Collin Baker, Christiane Fellbaum & Nancy Ide. 2012. The MASC word sense sentence corpus. In *Proceedings of the eighth language resources and evaluation conference (LREC)*, 3025–3030. Istanbul, Turkey. Project Website: http://www.anc.org/MASC/About.html.

Pearson, Matthew. 2005. The Malagasy Subject/Topic as an A-bar element. *Natural Language and Linguistic Theory* 24. 381–457.

Phillips, Colin. 2003. Linear order and constituency. *Linguistic Inquiry* 34. 37–90.

Phillips, Colin & Shevaun Lewis. 2009. Derivational order in syntax: Evidence and architectural consequences. To appear in *Directions in Derivations*, C. Chesi, ed., Elsevier.

Pilevar-Taher, Mohammad, Heshaam Faili & Abdol-Hamid Pilevar. 2011. TEP: Tehran English-Persian parallel corpus. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing*, vol. 6609 Lecture Notes in Computer Science, 68–79. Springer Berlin Heidelberg. doi:10.1007/978-3-642-19437-5_6. http://dx.doi.org/10.1007/978-3-642-19437-5_6.

Potsdam, Eric. 2003. Ellipsis identity and Malagasy sluicing. In Jason Merchant (ed.), *Sluicing: Cross-linguistic perspectives*, Amsterdam: John Benjamins.

Rohde, Douglas L. T. 2001–5. TGrep2: the next-generation search engine for parse trees. Department of Brain and Cognitive Science, MIT. http://tedlab.mit.edu/~dr/Tgrep2/.

Romero, Maribel. 1997. Recoverability conditions for sluicing. In Francis Corbin, Danièle Godard & Jean-Marie Marandin (eds.), *Empirical issues in formal syntax and semantics: Selected papers from the Colloque de syntaxe et de Sémantique de Paris*, 193–216. New York: Peter Lang.

Romero, Maribel. 1998. *Focus and reconstruction effects in wh-phrases*. Amherst: University of Massachusetts dissertation.

Rooth, Mats. 1992. Ellipsis redundancy and reduction redundancy. In Steve Berman & Arild Hestvik (eds.), *Proceedings from the Stuttgart ellipsis workshop*, vol. 340 (Arbeitspapiere des Sonderforschungsbereichs 9), 1–26. Heidelberg: Universität Stuttgart.

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

Sag, Ivan. 1976. *Deletion and logical form*: MIT dissertation.

Samko, Bern. 2013. A feature-driven movement analysis of english participle preposing. In R. E. Santana-LaBarge (ed.), *Proceedings of WCCFL 31*, Cascadilla Proceedings Project. To appear.

Samko, Bern. 2014. Verb phrase preposing as verum focus. Presented to the 88th annual meeting of the Linguistic Society of America, Minneapolis, Minnesota.

Schieber, Stuart, Fernanda Pereira & Mary Dalrymple. 1999. Interaction of scope and ellipsis. In Shalom Lappin & Elabbas Benmamoun (eds.), *Fragments: Studies in ellipsis and gapping*, 8–31. Oxford: Oxford University Press.

Schuyler, Tamara. 2001. Wh-movement out of the site of VP ellipsis. In Séamas Mac Bhloscaidh (ed.), *Syntax and semantics at santa cruz*, vol. 3, 1–20. Santa Cruz, California: Linguistics Research Center, University of California Santa Cruz.

Shahabi, Mitra & Jorge Baptista. 2012. A corpus-based translation study on English-Persian verb phrase ellipsis. *ICAME Journal* 36. 95–112.

Socher, Richard, John Bauer, Christopher D. Manning & Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL 2013*, http://nlp.stanford.edu/software/lex-parser.shtml.

Stainton, Robert J. 2006. Neither fragments nor ellipsis. In Ljiljana Progovac, Kate Paesani, Eugenia Casielles & Ellen Barton (eds.), *The syntax of nonsententials: Multidisciplinary perspectives* Linguistics Today Series, 93–116. Philadelphia and Amsterdam: John Benjamins.

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

Takahashi, Shoichi & Danny Fox. 2005. MaxElide and the re-binding problem. In Aaron Lawson (ed.), *SALT XV, proceedings from Semantics and Linguistic Theory XV*, 223–240. Ithaca, New York: Cornell University, CLC Publications.

Tanaka, Hidekazu. 2011a. Syntactic identity and ellipsis. *The Linguistic Review* 28. 79–110.

Tanaka, Hidekazu. 2011b. Voice mismatch and syntactic identity. *Linguistic Inquiry* 42. 470–490.

Williams, Edwin. 1977. Discourse and logical form. *Linguistic Inquiry* 8. 101–139.

Yáñez-Bouza, Nuria. 2011. ARCHER Past and Present. *ICAME Journal* 35. 205–236. Project website: http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/.

Yoshida, Masaya, Jiyeon Lee & Michael Walsh Dickey. 2013. The island (in)sensitivity of sluicing and sprouting. In Jon Sprouse & Norbert Hornstein (eds.), *Experimental syntax and island effects*, 360–376. Cambridge and New York: Cambridge University Press.