Prepositional phrase attachment ambiguities in declarative and interrogative contexts: Oral reading data

Tyler J. Peckenpaugh

2019-08-19

Contents

List of Tables

List of Figures

Abstract

Abstract is a work in progress. This paper reports a study on the effect of interrogativity on the oral reading of temporarily ambiguous prepositional phrases (PPs). Specifically, it looks at sentences ending in a of two PPs, where the first is interpretable as the goal argument of the preceding verb, and the status of the second (PP2 Status) is manipulated to either necessarily be the goal argument of that verb (Arg), forcing reanalysis, or not (Mod), allowing the original parse to stand. No evidence is found that interrogativity impacts the difficulty of understanding the Arg-type sentences, despite an intuitive decrease in difficulty when those sentences are presented in an interrogative context. A double-reading protocol is employed, where participants are asked to read a sentence first without preview (Reading 1), and then after unlimited preview (Reading 2).A robust effect of PP2 Status is found for the prosodic phrasing of the target sentences, and an effect of interrogativity on the study time between Readings, Inter-Reading Time (IRT), is reported.

Acknowledgements

TBD

About this draft

This represents the document that will be defended on August 29th. The goal is to deposit before September 16, after incorporating whatever revisions are requested.

Chapter 1

Introduction and background

This paper presents a study on human sentence processing, or parsing, and on the parsing of a particular sort of ambiguity. Parsing is assumed to be the projection of structure by a reader or listener over a string (which obviously lacks inherent structure). Following the sort of model of parsing put forward by, e.g., Kimball (1973), Frazier & Fodor (1978), and Frazier & Clifton (1996), this study assumes that parsing is done online and that most material must be incorporated into the structure being built as soon as it is encountered. This can lead to mis-parses, where the parser has guessed wrong about how to incorporate a given phrase with a temporarily ambiguous structure, and then encounters material that cannot be incorporated into the current structure. When this happens, the parser must reanalyze the material that had so far been processed, in order to come upa structure that can accomodate both the new and old material grammatically~~discard part or all of its pending structure and try again to incorporate the material encountered so far~~. This sort of parser crash is ~~often~~[JDF] called a garden path.

[1]: This paragraph was ¶1 chapter 1

~~As briefly mentioned earlier, ~~“~~G~~garden path effects” occur when a temporarily ambiguous sentence resolves in such a way that the structure initially preferred by

[2]: From here until next note, was introductory section of chapter 2 the parser is incompatible with how the sentence actually continues. These parsing errors have traditionally been attributed to structurally-focused parsing preferences (Frazier, 1979; Frazier & Fodor, 1978; Kimball, 1973) that ignore semantic content on the first pass. Frazier (1979) formulates several of these, including the following two which are widely accepted in one form or another:

(1) *Minimal attachment* Attach incoming material into the phrase-marker being constructed using the fewest nodes consistent with the well-formedness rules of the language under analysis (Frazier, 1979, p. 24)

(2) *Late closure* When possible, attach incoming material into the clause currently being parsed[JDF] (Frazier, 1979, p. 20)

Because these strategies ignore semantic and pragmatic plausibility and the parser typically does not know what material might be further on in the string, mis-parses at temporarily ambiguous regions can occur, resulting in garden paths. *Minimal Attachment* (MA)[JDF] is important to this study and will be revisited later on.

An example is the commonly studied garden path sentence, “The horse raced past the barn fell” (Bever, 1970). Here, the initial parse incorrectly assumes that the matrix subject is the unmodified NP the horse, per Minimal Attachment, and takes the matrix verb to be raced, as in the sentence, “The horse raced past the finish line.”

(3) The horse raced past the barn fell (Bever, 1970)

a) [$_{S}$ [$_{NP}$ The horse] [$_{VP}$ raced past the barn]] ??? [$_{VP}$ fell]

b) [$_{S}$ [$_{NP}$ The horse raced past the barn] [$_{VP}$ fell]]

An attempted parse resulting in structure (3 a) crashes, as it is not possible to incorporate the final word fell in a grammatical way. Reanalysis is required, with the grammatical parse being (3 b) where the matrix subject is *the horse raced past the barn*, a noun phrase (NP) containing a reduced relative clause *raced past the* *barn*. Thus *fell* can be incorporated as the matrix verb, with a structure comparable to, "The horse (that was) raced past the barn was hungry."

~~There is an ongoing debate in the literature about what parsing model best fits the empirical facts. This study follows [@frazier1996construal] in assuming that structure-first parsing strategies are at play, in addition to a primary vs. non-primary relation distinction that determines how immediately a phrase must be incorporated into a parse, allowing for some material to be incorporated later and thereby make use of additional information that is not available for immediate parsing decisions.~~

The study being reported is concerned with certain sentences that contain ~~such~~ a temporarily ambiguous prepositional phrase (PP1), followed by another ~~PP~~JDF (PP2) which causes the expected parse to crash~~interferes with what is assumed to be the default interpretation of PP1~~ Specifically, it is expected that PP1 in (4) will initially be interpreted as the goal of *cram, but that parse will fail when it is realized that PP2 cannot plausibly modify* drawer.

[3]: Back to what was chapter 1, ¶ 2

(4) He had planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ in~~to~~ his briefcase].

(5) He had planned to cram the paperwork [$_{PP_1}$ in the drawer ~~]~~ [$_{PP_2}$ of his filing cabinet]].

[4]: examples moved up slightly

This contrasts with the similar sentence in (5) where PP2 can plausibly modify *drawer* and so the parse where *in the drawer* is the goal argument is accepted and *of his filing cabinet* is incoporated as a modifier within PP1.~~: as an argument of the verb. The reasons for that assumption are discussed in Section refmech, but for now it suffices to understand that (@cramiGPy) represents a difficult to comprehend sentence (a garden path), while (@cramiGPn) presents no such difficulty.~~

((examples were here))

~~Before the details of the current research can be outlined, it is first necessary to explain some of the terms and mechanisms involved. This chapter is concerned with doing so.~~

\begin{note} Sections 1.1 and 1.2 are redundant and have been removed. They are incorporated into the above. \end

1.1 Motivations for the current

[5]: Adapted from what was §1.3

The current study was initially motivated by an observation discovered by Janet Dean Fodor and Dianne Bradley, and originally reported in @qp2. That observation is that a garden path sentence like (4) repeated here as (6) is, for whatever reason, not as difficult to process when presented as an interrogative, as in (7), rather than a declarative.~~These attachment ambiguities which are difficult to parse in the declarative, appear to be less difficult to parse when presented in polar interrogative context. Consider again the somewhat difficult to process sentence in (@cramiGPy), repeated below as (@dec) for convenience.~~

(6) He had planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ in~~to~~ his briefcase].

(7) Had he planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ in~~to~~ his briefcase]?

Peckenpaugh (2016) attempted to find a behavioral correlate of this intuition by looking at variation in reading time of sentence like (6) and (7), but the results were inconclusive. The current study continues that line of research by looking at similar sentences, while also attempting to control additional factors that may have led to Peckenpaugh (2016)'s inconclusive results. One such concern is that (6) relies on the implausibility of a drawer within a briefcase, i.e., real world knowledge, in order to disambiguated the appropriate attachment sites for PP1 and PP2. Real world knowledge and beliefs of what is or is not plausible varies between speakers and may not always be reliable as a trigger for reanalysis. The current study makes use of carefully constructed sentences (the criteria by which they were constructed is detailed in Section 2.1) which do not rely on plausibility or pragmatics to disambiguated the PP attachment sites, but instead make use of syntactic disambiguation by including a PP2 that cannot grammatically be incorporated into PP1, as in (8).

(8) He had planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ *into* his briefcase].

The change from pragmatically disambiguation to syntactic disambiguation creates an important distinction between the original observation in Peckenpaugh (2016) and what the current study is considering. It cannot be assumed that the intuition about (6) vs. (7) necessarily extends to cases like (8), and so one of the questions the current study asks is whether the intuition can be shown to extend to syntactically disambiguated cases that are similar to the pragmatically disambiguated cases for which the observation was first made. In order to keep clear this fine but important difference, a convention is adopted throughout this document where the intuited amelioration of the garden path effect in pragmatically disambiguated cases is referred to as the 2016 intuition, while the possiblity that the intuition extends to syntactically disambiguated cases is referred to as the current hypothesis.

(9) *The 2016 intuition:* Certain pragmatically disambiguated prepositional phrase (PP) attachment ambiguities which are difficult to parse in the declarative are less difficult to parse when presented as yes-no interrogatives (e.g., Jed crammed the newspapers under the sofa in the trashcan).

[6]: these are new to this chapter, but are taken from the discussion chapter, where we discussed them on 8/16

(10) *The current hypothesis:* The intuition may be extensible to PP attachment ambiguities that are syntactically disambiguated in addition to those that are pragmatically disambiguated (e.g., Had he planned to cram the paperwork [$_{\mathrm{PP1}}$ in the drawer] [$_{\mathrm{PP2}}$ into his briefcase]?).

~~While both sentences contain a temporary ambiguity for the attachment of PPs occurring in string-linear sequence at the end of an utterance, it is more likely in (@intr) that a listener or reader will come away with a plausible interpretation on the first try. This study explores the factors involved in why that is, and seeks to uncover a behavioral correlate of this intuition. If the existence and mechanics of this effect can be pinned down, it will lend insight into what information the parser has access to when making decisions, or when repairing broken parses. This initial observation reveals a robust program of research with many interwoven questions.~~

In addition to exploring the possible extensibility of the 2016 intuition, the current study is interested in~~One might also ask~~ what property or properties of polar questions might lead to an easier parsing of garden paths, or at least to the perception that they are easier to parse when compared to~~as polar questions than as~~ declaratives. Because there are minimal differences between the polar question and declarative version of declarative versions of these sentences~~a given sentence~~, it is fairly easy~~reasonable~~ to assume that the cause lies in one of two domains: the prosodic changes triggered by the use of question intonation, or the pragmatic and semantic properties that are not shared across the versions.

An obvious reflex of the former possibility is another question: how are the various versions of these sentences actually pronounced, prosodically? ~~This question is deceptively difficult to answer;~~ ~~T~~the reported study seeks to answer it, but while the recordings collected provide some insight, more work is likely needed to satisfactorily provide an answer.

Likewise, the latter possibility leads one to wonder: what are the semantic and pragmatic differences between a polar question and its declarative counterpart? This question can be approach in a more theoretical way, and has been to some degree in the literature. That said, it remains to be determined how those properties could lead to an easier parsing process, and whether or not a satisfactory explanation for the intuition at large can be pulled from these differences.

1.2 Structural overview of the ambiguity relevant to this study

The 2016 intuition and the current hypothesis are both concerned with a temporarily ambiguous sequence of PPs at the end of a sentence. This section will discuss what the possible attachment sites for those PPs are and which structures ultimately do and do not work.

[7]: the following is adapted from what was §2.1

The example in (11) shows a pragmatically disambiguated sentence with an argument-PP2. The initial parse (a) is implausible, resulting in structural reanalysis to the preferable parse in (b).

(11) Jed crammed the newspapers under the sofa in the wastebasket~~trashcan~~.

a) # ... [$_{VP}$ crammed [$_{NP}$ the newspapers]
[$_{PP1}$ under [$_{NP}$ the sofa [$_{PP2}$ in the wastebasket]]]

b) ✓ ... [$_{VP}$ crammed [$_{NP}$ the newspapers [$_{PP1}$ under [$_{NP}$ the sofa]] [$_{PP2}$ in the wastebasket~~trashcan~~]] *"#" indicates a structure with an implausible reading*

The initial parse is expected to be (a) because of~~In parsing (@aa a), there is~~ a ~~fairly strong~~ bias (due to *Minimal Attachment*, or some variation thereof), which favors a structure where the first PP attaches into the verb phrase (VP) as an argument of the verb, i.e., [$_{VP}$ V NP PP1], which leaves nowhere for the second PP to attach but as a modifier of the noun phrase (NP) inside PP1 ([$_{\text{PP1}}$ *under* [$_{\text{NP}}$ *the sofa* [$_{\text{PP2}}$ *in the trashcan*]]]). This initial parse (11 a) is pragmatically implausible, as one does not generally find sofas inside wastebaskets. Structural[JDF] r~~R~~eanalysis is required to bring about the correct parse (11 b), where PP1 attaches as an NP modifier of the direct object and so allows PP2 to attach as a VP argument, resulting in a structure such as [$_{\text{VP}}$ V [$_{\text{NP}}$ N PP1] PP2], i.e., where it is the newspapers under the sofa that are being crammed in the trashcan.

It is important to be clear on how the sentences of concerned are structured. As a matter of terminology, the current study categorizes the sentences being discussed into two groups based on the status of the PP2 they contain: (a) cases where PP2 is an argument are Arg-type sentences, and (b) cases where PP2 is a modifier are Mod-type sentences. In practice, Arg sentences are garden paths, because PP2 must fill the goal role that PP1 is expected to have filled as just discussed. Mod sentences are not garden paths, because PP2 can modify the NP within PP1 and become part of the goal, and therefore need not disrupt PP1 from being the goal.

(12) **PP2 Status**

[8]: this example has been moved up

(a) *PP2 Argument* (Arg)

He had planned to cram the paperwork [$_{\text{PP1}}$ in the drawer] [$_{\text{PP2}}$ into his briefcase].

(b) *PP2 Modifier* (Mod)

He had planned to cram the paperwork [$_{\text{PP1}}$ in the drawer [$_{\text{PP2}}$ of his filing cabinet]].

In order for the differing parsing process for (12 a) and (12 b) to be explained by a strictly structurally based model of parsing, certain assumptions ~~would~~ have to be made about the syntax. A simple way to get the explanation to work is to assume that all arguments of a verb are syntactic sisters to the verb, resulting in a three-way branching VP for ditransitive verbs. In this case, in order to avoid postulating extra nodes that would be required for PP1 to be a modifier, *Minimal Attachment* dictates that PP1 should be assumed to fill the argument slot. This is not how modern syntactic theory assumes the structure looks, as three-way branching is proscribed. Figures 1.1 and 1.2 show the presumed ~~modern~~ structure[1] of these sentence~~, graphically~~.

[1]Note that when the internal structure of an NP is not relevant (no PP is within it) it is not drawn, i.e. [NP newspapers] is shorthand for [NP [N' [N newspapers]]].

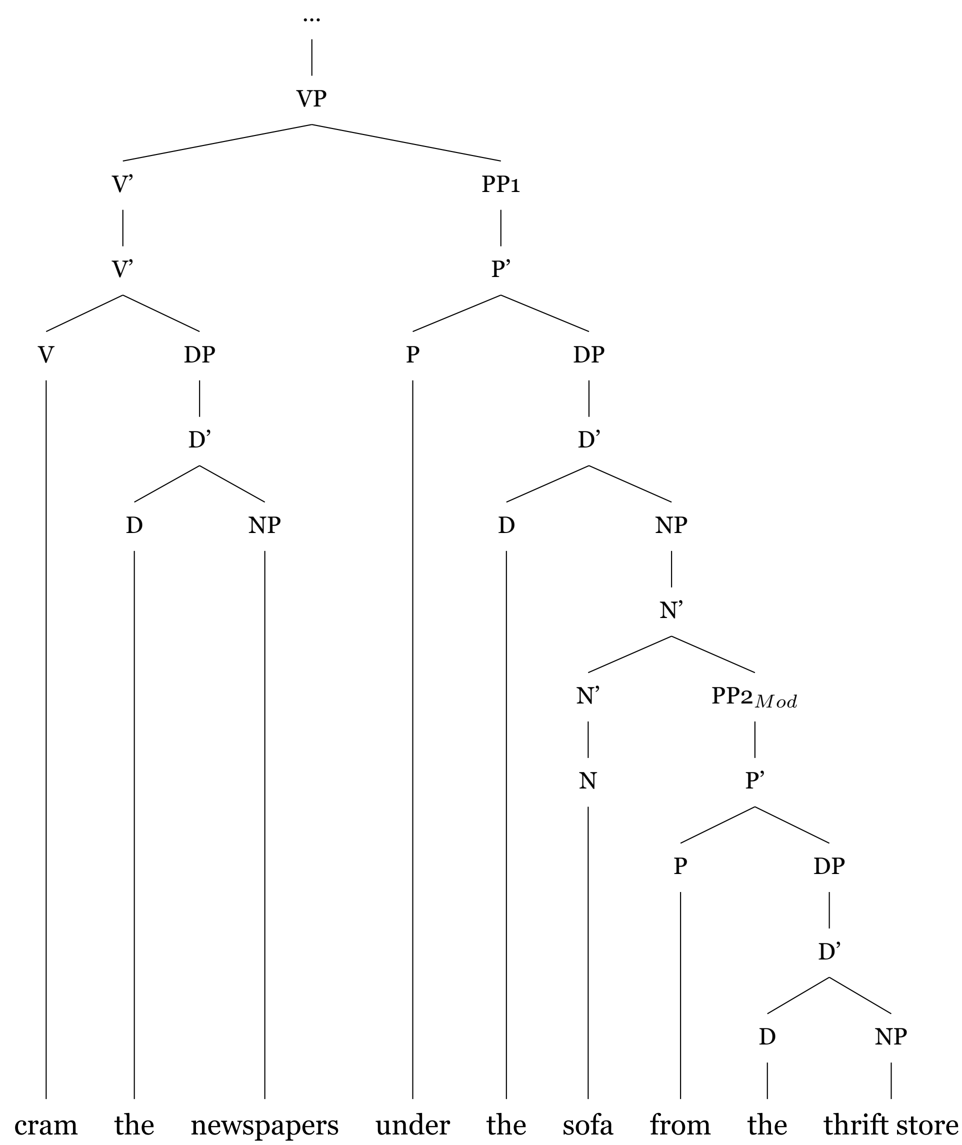

Figure 1.1: Syntactic tree of an illustrative example sentence with an ambiguous PP1 and a modifier-PP2 (Mod).

~~Note that~~T~~t~~he most pronounced structural difference between the two structures is that in the sentence with a modifier-PP2 (Mod)~~Modifier~~ the major disjuncture ~~, i.e., change in branching direction,~~ comes fairly~~much~~ earl y~~ier~~, just after the object NP *the newspapers*. For the ~~Arg~~ sentences with argument-PP2s (Arg), the major disjuncture~~branch direction change~~ is ~~much~~ later, just after PP1 (*under the sofa*).

~~This paper is not interested in the particularities of syntactic theory, and it also is~~

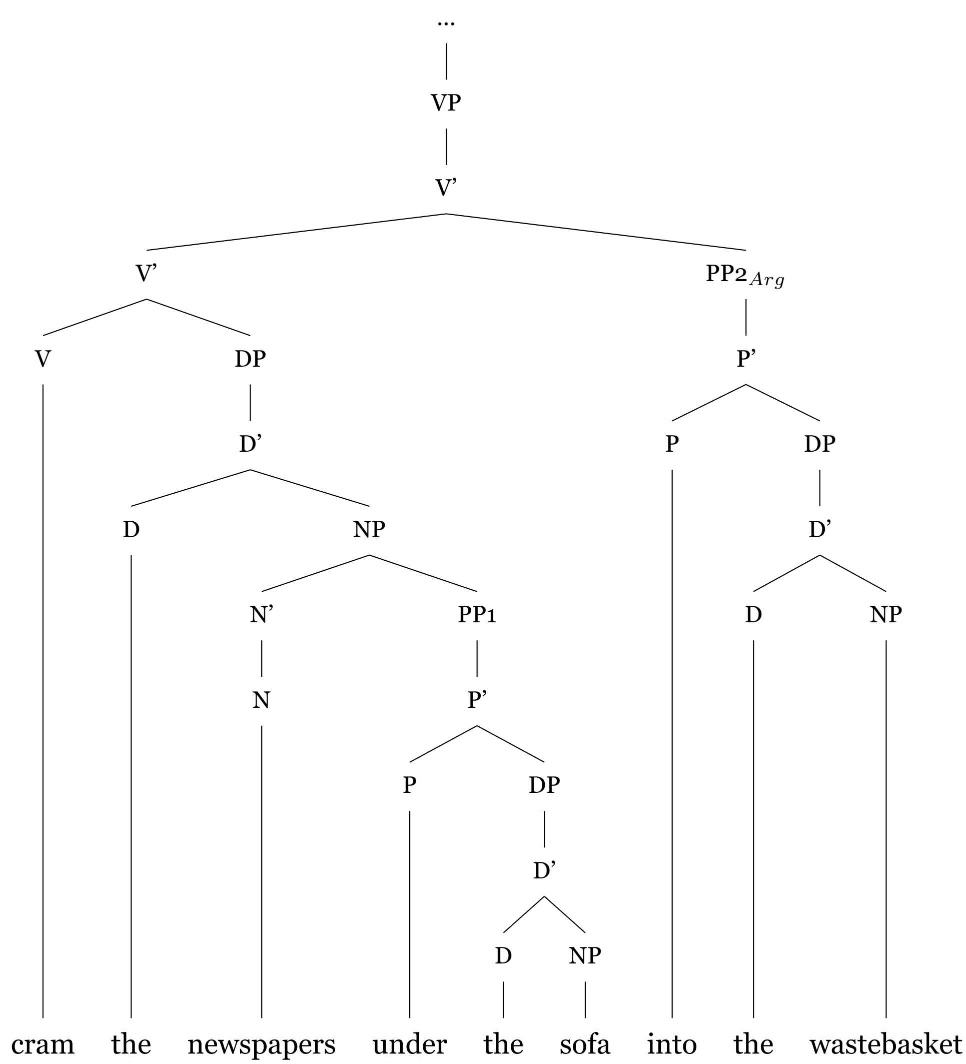

Figure 1.2: Syntactic tree of an illustrative example sentence with an ambiguous PP1 and an argument-PP2 (Arg).

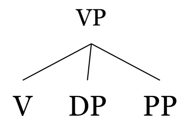

Figure 1.3: Illustrative syntactic tree of a ternary-branching VP.

~~not necessary to rely on *Minimal Attachment* to make the necessary distinction, though it very likely does play a role. Instead, we can focus on distinction added to parsing theory in Construal [@frazier1996construal]: that of primary vs. non-primary relations.~~

~~This study is focused on the impact of Speech Act, i.e., where a sentence is interrogative (Q) or declarative (D), in a particular sort of garden path. Specifically, it is concerned with garden path sentences containing a temporary ambiguity that centers on the attachment of two prepositional phrases (PPs) occurring in string-linear sequence at the end of an sentence, e.g., “When we saw her, the nanny had seated the cranky little boy [PP1 on the swing] [PP2 in his stroller].”~~

It bears mentioning~~Note~~ that *Minimal Attachment* as defined by (Frazier, 1979) is somewhat at odds with recent developments in syntactic theory, e.g., obligatory binary branching (cf. Chomsky, 2014, p. 62). As originally postulated, *Minimal Attachment* relies on~~assumes that~~ a verb with multiple internal arguments incorporat~~ing~~es each of those arguments as a sister (i.e., a ternary branching structure as in 1.3).

[9]: new figure: ternary tree

With current theories of syntax where binary branching is obligatory, two XPs (NP and PP) cannot both be syntactic sisters of the verb, and the structures are assumed to be as shown in the trees presented earlier (Figures 1.1 and 1.2) so it becomes less clear that the VP attachment site for PP1 actually creates fewer nodes than the lower NP attachment site. Nonetheless, the preference for VP attachment in these kinds of sentences is there, be it due to Minimal Attachment, a preference for arguments over non-arguments, or something else, as evidenced by experimental data from e.g., Rayner, Carlson, & Frazier (1983) and Clifton, Speer, & Abney (1991).

~~This study is focused on the impact of Speech Act (interrogative vs. declarative) and its interaction with a trailing sequence of prepositional phrases (PPs), where the second is of two possible types. The contrasting types of PP2 shown in (@pp2t) are (a) a PP2 which must be an argument, and (b) one which can be a modifier.~~

~~PP1, *in the drawer*, is the same in both (@pp2t a) and (@pp2t b), and is ambiguous on first encounter, as it could modify the paperwork or it could be the goal of cram. In (@pp2t a), however, PP2 must ultimately be interpreted as the goal of the verb *cram* because *into his briefcase* cannot modify *the drawer*. Because *cram* only accepts one goal, this means that PP1 in (@pp2t a) has to end up as a modifier of *the paperwork*. In (@pp2t b), on the other hand, PP2 *of his filing cabinet* can (in this case, must) modify *the drawer*, and so in the drawer can and does end up as the goal of cram. The difference in PP2 Status between (@pp2t a) and (@pp2t b) results in different structures, which I argue are reached by different parsing mechanisms. Namely, (@pp2t a) should, by hypothesis, result in a parse which initially incorporates PP1 as the goal argument of *cram* but then fails and triggers reanalysis when PP2 is encountered. Conversely, (@pp2t b) should by hypothesis allow a straightforward parse where PP1 is initially and ultimately slotted in as the goal of *cram*, since PP2 poses no issue when interpreted as a modifier of *the drawer*. This means (@pp2t b) should not trigger reanalysis. Where (@pp2t a) is a so-called garden path sentence, (@pp2t b) is not. In what follows, I will use the term ”argument attachment of PP2” to mean the garden path case, i.e., a sentence that is presumed to require reanalysis, and ”modifier attachment of PP2” to mean the straightforwardly parsed case where PP1 is the goal argument.~~

Rather than worry about~~This paper is not interested in~~ the particularities of syntactic theory~~,~~ and it also is not necessary to rely on~~the exact formulation of~~ *Minimal Attachment*, an appeal can be made to~~make the necessary distinction, though it very likely does play a role. Instead, we can focus on~~ a distinction added to parsing theory in *Construal* (Frazier & Clifton, 1996): that of primary vs. non-primary relations. ~~The impetus behind adding this T~~this additional machinery ~~to the theory of parsing~~ is independently motivated: while structure-first decision making seems to hold for the parsing of many structures, there are some that seem to follow different rules~~flout them~~. *Construal* illustrates this by way of relative clause (RC) attachment in constructions like (13).

[10]: collapsed two ¶into one

(13) [$_{NP1}$ The daughter$_i$] of [$_{NP2}$ the colonel$_j$] [$_{RC}$ who$_{i/j}$ was standing on the balcony] ...

The RC in (13) can modify either NP1, the daughter, or NP2 the colonel. A structure-first parsing system, together with the widely agreed upon structural parsing strategy *Late Closure*, predict~~would be expected to manifest as~~ a consistent preference for ~~the~~JDF local attachment of the RC in (13), i.e., the structure where the RC modifies NP2. Instead, what ~~Frazier and Clifton describe, based on~~ a number of empiricalJDF studies (e.g., Clifton, 1988; Cuetos & Mitchell, 1988) find is a pattern where the preferred structure depends on the relationship between NP1 and NP2. ~~They~~Frazier & Clifton (1996) describe five categories of relationship, and a gradient of preferred RC attachment, from NP1 preference to NP2 preference.

(14) **RC Attachment by NP1-NP2 relation** (Frazier & Clifton, 1996)

a) *Material*

The table of wood [$_{RC}$ that was from Galicia]

b) *Quantity*

The glass of wine [$_{RC}$ you liked]

c) *Relational* (friend, enemy, son, and other argument taking NPs, e.g., picture-NPs)

The son of the woman [$_{RC}$ that was dying]

d) *Possessive*

The car of the company [$_{RC}$ that was falling apart]

e) *Non-accompaniment* with

The girl with the hat [$_{RC}$ that looked funny]

Frazier & Clifton (1996) cite studies showing~~report~~ that (14 a-b) type configurations favor NP1 RC attachment, (14 e) type configurations favor NP2 RC attachment, while (14 c-d) are intermediate. They argue that this gradient cannot be readily explained by structural parsing, and instead make use of a mechanism they call ~~structural~~ association. RCs, rather than being immediately slotted into a tree in a specific way, are associated with a thematic domain, i.e., the maximal projection of whatever lexical item last assigned theta-roles, together with associated functional projections; in the case of the examples in (14), the last theta assigner is NP2, and its domain extends up to the DP that contains NP1. This is a later~~looser~~ parsing decision that allows the syntactic structure to be decided ~~on later,~~ after semantic information becomes available: the RC can ultimately modify whichever member of the thematic domain is appropriate.

The crucial issue that distinguishes cases where structural association vs. structural parsing is appropriate is the idea of primary vs. non-primary relations. Frazier and Clifton formalize this distinction as (15).

(15) "Primary phrases and relations include

a) The subject and main predicate of any (+ or -) finite clause

b) Complements and obligatory constituents of primary phrases" (Frazier & Clifton, 1996, p. 41)

[JDF 1]: citation moved and quotation marks added to show this is a direct quote

RC attachment undergoes association because the relationship between a modifier and whatever is modified is a non-primary relation, and a relative clause is by definition a modifier and not an argument. Circling back to the PP-attachment that this study is concerned with, the argument vs. modifier distinction is precisely what distinguishes the two possible statuses of PP2 shown and illustrated[JDF] in (12) and repeated here as (16).

(16) **PP2 types**

(a) *PP2 Argument* (Arg) He had planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase].

(b) *PP2 Modifier* (Mod) He had planned to cram the paperwork [$_{PP_1}$ in the drawer [$_{PP_2}$ of his filing cabinet]].

Without locking down the exact syntactic structures that (16) represents, we can nonetheless say that the parser would seek to immediately incorporate PP1 into the tree in both cases. The infinitival (-finite) clause headed by *cram* is a primary phrase, and so its obligatory constituents hold primary relationships with *cram*. *Cram* takes an obligatory goal argument, so the parser cannot wait for semantic information to inform its association, it must make its best guess based on the principles of structural parsing, and attach it as an argument, as that property is what is forcing the immediate decision to be made. In the case of (16 a), when PP2 is encountered reanalysis of the PP1 attachment[JDF] will be required, due to the fact that PP2 must be the goal of *cram* and therefore takes the syntactic position that PP1 had been filling. In the case of (16 b), no reanalysis is necessary because.

[11]: this bit has been reworked

1.3 Interrogativity

This section returns to the questions raised by the 2016 intuition discussed in Section 1.1 about why certain interrogative garden paths of the sort just discussed might appear to be easier to parse than similar declarative ones.

(17) He had planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ in~~to~~ his briefcase].

(18) Had he planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ in~~to~~ his briefcase]?

Specifically,~~The question that must be asked, then, is~~ what exactly differs between (17) and (18)? Syntactically, very little: the position of the subject he and the auxiliary had have been reversed.

\deleted{Semantically, or perhaps it is better to say pragmatically, there are a number of differences, which are~~I will~~ \deleted[id=JDF]{perfunctorily{} discussed briefly here'[JDF]. ~~The details may not be quite rightwrite, as the focus of this paper lies elsewhere, but it's important to be aware of the general ideas presented.~~

The semantic, or perhaps more accurately pragmatic, differences between (17) and (18) lie with the presuppositions the sentences carry with them, and with the placement of focus. The declarative in (17) has few presuppositions beyond the existence of the actors and objects involved (the referents of *he*, *paperwork*, *drawer* and *briefcase*), and that these actors and objects can be involved in *cramming*. The presuppositions of (18) are different froma~~ super set of~~ those of (17): a yes/no question additionally presupposes that the listener knows the answer to the question, for one, and also may not presuppose the referent of a focused element. The use of focus introduces alternative possibilities for what might be filling a given role, i.e., in (@foc5) where focus is on *paperwork*, the question is whether it was paperwork that was crammed into the drawer, and so the existence of paperwork is not necessary if the answer to the question turns out to be "no." Further presuppositions might exist, depending on where the focus lies within the sentence.

Focus in a declarative like (17) is typically broad, meaning no element is having attention called to it. A polar question like (18), however, will typically receive narrow focus on one element, so that when uttered, one element is more prominent than the others. The focused element becomes the part of the sentence that the question is about. Focus can fall on any of the lexical or referential elements (subject, matrix verb, infinitival verb, object NP, or the NP of either PP1 or PP2) of the sentence, or the auxiliary verb. Bold facing indicates verbal emphasis in order to locate focus.

(19) **Had** he planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase]?

(20) Had **he** planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase]?

(21) Had he **planned** to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase]?

(22) Had he planned to **cram** the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase]?

(23) Had he planned to cram the **paperwork** [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his briefcase]?

(24) Had he planned to cram the paperwork [$_{PP_1}$ in the **drawer**] [$_{PP_2}$ into his briefcase]?

(25) Had he planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into **his** briefcase]?

(26) Had he planned to cram the paperwork [$_{PP_1}$ in the drawer] [$_{PP_2}$ into his **briefcase**]?

In (19), with focus on the auxiliary, the question is about the entire proposition, and whether or not it is true. In this case, there are not any additional presuppositions when compared to the declarative counterpart of the sentence. In (20), with focus on *he*, the question is asking about whether the referent of *he* is the actor who performed the action described; in this case, the entire predicate is presupposed: someone *had planned to cram the paper in the drawer into his briefcase*, but was it *him*? Skipping ahead to (24), with focus on *drawer*, the question is instead about which *paperwork* this is all happening to: *the paperwork in the drawer*, or some other stack of paperwork? In this case, it is presupposed that the referent of *he* was the one who *had planned to cram some paperwork into his briefcase*, and only the exact referent of *the paperwork* is not presupposed. For each other location of focus, the presuppositional content is similarly complementary to whichever element is focused and therefore being asked about.

This set of pragmatic differences between (17) and (18) might very well be the source of the 2016 intuition that (18) is easier to comprehend than (17), though the details of how that would work are not entirely clear, ~~b~~ut that is not the only possibility. Another significant difference between the two Speech Acts is the prosody and intonational melody. While dialects of English differ, there is typically a difference in melody between a declarative and question, and in many American English dialects, the interrogative is pronounced with a final rise, while the declarative exhibits just a series of down-steps. This difference is the one that the current study explores, to see if it can readily explain the intuitive difference in processing difficulty.

1.4 Prosody of questions vs. declaratives

In pursuing the possibility that it is the intonation and prosody of polar interrogatives which creates the 2016 intuition~~intuitive contrast~~ that motivated this study ~~investigates~~, we must consider what question intonation actually sounds like. It is generally agreed that in American English, the intonation of a polar ~~(yes/no)~~ question has the property of a final rise. Indeed, this has been confirmed in corpus studies such as Hedberg, Sosa, & Görgülü (2017) who found that 79.8% of the 410 American English yes/no questions in their study (ten-minute phone conversations from the CallHome Corpus of American English and the Fisher English Corpus) had a "low-rise nuclear contour" (L*H-H%, L*H-↑H%, or L*L-H%)[2]. In~~To briefly explain~~JDF their ToBI notation, a tone T is either L for low or H for high; T* is anchored to the stressed syllable, and T- and T% are boundary tones (intermediate phrase boundary and intonational phrase boundary respectively). See, e.g., *Guidelines for ToBI labeling* (Beckman & Ayers, 1997) for a more thorough explanation of ToBI. An additional 10.7% of the Hedberg et al. (2017) data had a "high-rise nuclear contour" (the authors categorizes the following tunes as "high-rise nuclear contours:" H*H-H%, or !H*L-L%). That leaves only 9.5% spread across the other 5 categories (High-fall, Rise-fall, Low-fall, Fall-rise, and Level). Only 5.6% of the data showed a falling contour. According to the authors' analyses, these contours occur on the final main stress of a sentence and thereafter. In the case of the types of sentences examined in the current study, that would result in a rising contour on the head noun of the final PP as inJDF (27).

(27) Did Jed cram the newspapers under the sofa in the [$_{\text{L*H-H\%}}$ guestroom].

The need to prepare for that finalJDF rising tone might make a prosodic break before the PP more likely, and thus ease reanalysis or even encourage a different prosodic chunking which might encourage argument attachment. This is possibility is revisted and more fully explained in Section **??**.

~~A brief informal surveyinvestigation of the wave forms of recordings of several native speakers of American English found that most speakers maintain low tones on prior stresses, although some had a H tone on the subject noun. It also varied between speakers and between sentences~~

[12]: this was a pilot study I conducted with friends and family

[2]Hedberg et al. (2017) use ↑ to indicate an up-step, which is not standardly transcribed with ToBI.

~~as to whether there is a prosodic boundary (marked by a low tone and/or pause) immediately before the rise (after PP1) or not.~~

The prosodic structures found in the data collected for the current study are discussed in 3.2.

1.5 Can prosody affect parsing?

A number of studies have shown that in listening to speech, prosodic cues appear to help reduce the frequency with which incorrect parsing (i.e., a garden path) occurs. For example, Kjelgaard & Speer (1999) conducted a study using digital manipulation of recorded speech to create three versions of sentences containing a ~~garden path~~JDF temporary ambiguity ~~(discussed above)~~ which could result in a garden pathJDF. They recorded speakers saying sentences with natural prosody, such as the following pair (not bracketed in presentation to the participants)JDF:

(28) [When Roger leaves] the house is dark. (Early closure)

(29) [When Roger leaves the house] it's dark. (Late closure)

They then cross-spliced these together to make several versions. One version had prosodic cues which cooperated with the intended reading of the sentence; another attempted to have "neutral" prosody; and the third used intentionally misleading prosody. The initial fragment of each was then presented to participants (the portion from the beginning of the sentence to the word *house* in (28 - 29) and they were asked to agree or disagree with whether a visually presented word, either *is* or *it's* was likely to be the next word in the sentence. Participants gave more accurate and speedier judgments when the prosodic cues lined up with the correct parsing. The results of this study, as well as a growing body of literature, suggest that that prosodic information can ~~(or perhaps must)~~ be used by the parser in making processing decisions.

Consider the analysis by[JDF] Fodor (2002) ~~analysis~~[JDF] of relative clause attachment preferences in English[JDF], an example of the differing RC attachment preferences across languages pointed out by Frazier & Clifton (1996) and discussed in Section 1.2. This concerns sentences such as (30):

(30) Someone shot the servant$_{N1}$ of the actress$_{N2}$ [$_{RC}$ who was on the balcony].

The relative clause (RC) *who was on the balcony* can attach either locally (~~low,~~ modifying N2), making it *the actress* who was *on the balcony*, or non-locally (~~higher up (non-locally,~~ modifying N1), so that it is~~we understand~~ *the servant* ~~to be the one~~ who was *on the balcony*. While *Late Closure* predicts local/low attachment i~~I~~n these sorts of sentences, Cuetos & Mitchell (1988) found a 60% preference for low attachment in English speakers, but only a 40% preference for low attachment in Spanish speakers. In apparent violation of the general preference for local attachment, some languages, like French and Spanish (and Russian, but not Romanian or Brazilian Portuguese, so this is not a general feature of Romance languages), prefer to attach relative clauses higher, while others more often obey Late Closure (e.g., Swedish, Egyptian Arabic, and English). This non-local preference is weakened in cases where the ambiguous RC is short (one prosodic word). Fodor (2002) asserts that these tendencies exist both[JDF] in ~~both~~[JDF] listening to spoken words (under conditions where a particular parse is not favored by the explicit prosody) and in silent reading.

Fodor notes that other researchers have shown ~~the presence and absence of prosodic breaks to influence parsing decisions, and specifically~~[JDF] that the presence of a prosodic break before the RC in sentences like (30) encourages high attachment. Fodor leverages this in order to explain the difference in RC attachment site tendency between languages. She argues that the phenomenon can be neatly account for by linking attachment site preference to the likelihood of a prosodic break before the RC. This difference in prosodic tendency, in turn, can be explained using a constraints-based approach. Consider Selkirk's (1986) alignment constraints:

(31) Align(αCat, E; βCat, E)

a. Align (GCat, E; PCat, E)

b. Align (PCat, E; GCat, E)

c. Align (PCat, E; PCat, E) *GCat ranges over morphological and syntactic categories; PCat ranges over prosodic categories; E = Right or Left (Selkirk, 1986, p. 6)*

Truckenbrodt (1999) provides a prose-based formalization of the same idea. ~~He describes what I will call *Align R * which can be easily generalized to described what I will call *Align L *, the same constraint except that it calls for aligning phrases at their left edges rather than their right edges.~~

(32) **Align-XP/R**

For each XP there is a PP such that the right edge of the XP coincides with the right edge of the PP, where XP is a maximal projection and PP is a Phonological Phrase. This constraint represents the end based mapping assumption for Major Phonological Phrases in English, whose right end is supposed to align with the right end of Maximal Projections (Truckenbrodt, 1999, p. 223).

Essentially, Selkirk (1986) argues that relative ranking of alignment constraints for the left edge of phrases (*Align-XP/L*) with those for the right edge of phrases (*Align-XP/R*) can impact the distribution of prosodic breaks. These alignment constraints dictate that the edges of prosodic units (and thus the location of prosodic breaks) should align with the edges of syntactic constituents. Because the prosodic break that encourages high attachment is one which aligns with the left edge of the RC in examples like[JDF] (30), postulating that *Align-XP/L* is ranked above *Align-XP/R* in languages like French that prefer high attachment can account for that preference (remember that a prosodic break in that place has been shown to encourage a high attachment interpretation). In languages where low RC[JDF] attachment is preferred, we can assume that *Align-XP/R* is ranked higher, and thus a prosodic break is more likely to occur after the RC than before.

The same sort of argument can explain the difference in tendency between long and short RCs. Consider Selkirk's (2011) *BinMin* defined below.

(33) **BinMin(ϕ)**

A ϕ (phonological phrase) must consist of at least two ω (phonological words).

If we assume, in Optimality Theoretic (Prince & Smolensky, 1993) terms, that a constraint like *BinMin* is ranked above $Align_L$, then it seems quite reasonable to assume that a prosodic break before a short RC (which would encourage high attachment) is much less likely than before a long RC. That is, when the RC is short, its left edge is prevented from aligning with the beginning of a prosodic phrase (it violates $Align_L$) by the higher ranked BinMin. Longer RCs can have their left edge align with the start of a prosodic phrase, and thus can have the high-attachment encouraging prosodic break.

1.6 Predictions for the current study

This study is concerned with a number of issues. First: is attachment in any way encoded in the speech signal? I hypothesize, following e.g., Schafer, Speer, Warren, & White (2000), that we can use prosody to diagnose attachment site. Consider ~~Assume~~[JDF] the ~~following~~ basic configuration in Figures 1.4 and 1.5.

[13]: example 35 replaced by figures

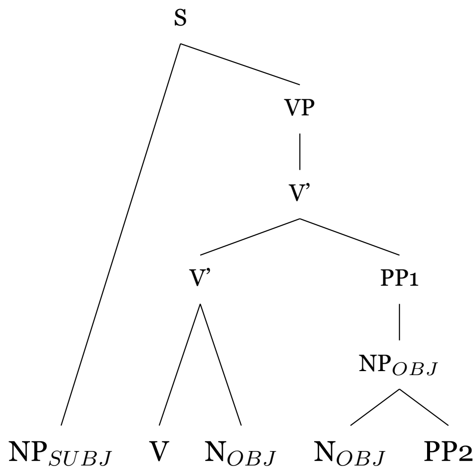

Figure 1.4: Illustrative syntactic tree of the basic configuration for Mod cases.

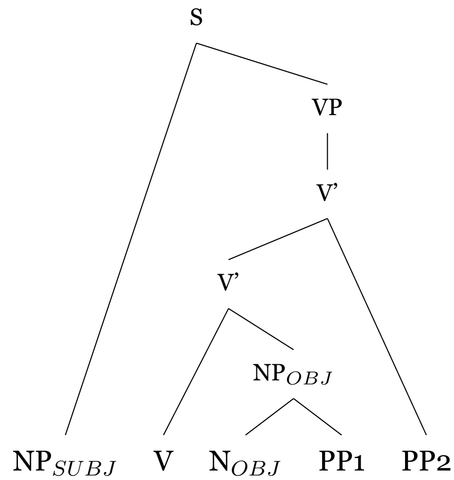

Figure 1.5: Illustrative syntactic tree of the basic configuration for Arg cases.

I suggest~~The current study proposes~~ that argument attachment of PP2 will be marked by a prosodic boundary between PP1 and PP2 (for discussion of what constitutes a prosodic boundary, see e.g., Streeter (1978) and Salverda, Dahan, & McQueen (2003)). Modifier attachment of PP2, on the other hand, will lack any substantial boundary marking in that position; instead, a break after the object is expected.

(34) *Hypothesis 1* Argument attachment of PP2 is marked by a dominant prosodic break between PP1 and PP2.

The second issue: what factors impact immediate on-line parsing, and what factors only affect later, post-parse considerations? To address this, the study will employ the double reading paradigm of Fodor, Macaulay, Ronkos, Callahan, & Peckenpaugh (2019) (more on double reading in the methods section). For example, if first-pass parsing ignores semantic information, then implausible parses should be more frequent in Reading 1 of a sentence with argument-PP2 than in a second reading of the same sentence.

(35) *Hypothesis 2* A first reading of a sentence where PP2 is a goal argument will exhibit less natural prosody (more hesitation at and within the PP2 region) than:

- A first reading of a sentence where PP2 is a modifier
- A second reading of a sentence where PP2 is a goal argument

Hypothesis 2 and *3* together make a third prediction: readers should struggle more on the cold reading of a GP sentence to obtain a plausible structure, and thus the appropriate prosody, than on a previewed reading.

(36) *Hypothesis 3* A first reading of a sentence with an argument-PP2 will more often be produced with prosodic structure that represents an implausible or ungrammatical parse of the string (PP2 incorrectly attached as a modifier), whereas a second reading sentence will more often be pronounced with the prosodic structure that represents the intended parse (argument attachment of PP2).

Note that *hypothesis 3* can't be applied in cases where the reader fails to successfully and fluently produce the sentence.

Finally, the question of whether the 2016 intuition~~Finally, I investigate the intuition~~ described in Section 1.1 can be shown to extend to the syntactically disambiguated sentences used in this study is returned to with *hypothesis 4*~~originally discovered by Dr. Janet Dean Fodor and Dr. Dianne Bradley: that these sentences are not as difficult to parse when encountered in interrogative, as opposed to declarative, context~~.

(37) *Hypothesis 4* Reading 1 of a declarative sentence with an argument-PP2 will exhibit less natural prosody (more hesitation at and after the disambiguating region) and be more likely to be produced with prosodic structure that represents an implausible or ungrammatical parse of the string than a

Reading 1 of an interrogative sentence with an argument-PP2.

These hypothesis are returned to in Chapter 3. The findings presented in the current ~~this~~ study do not successfully settle all of the issues raised~~answer all of the questions asked~~JDF here, but it is hoped that they~~do~~JDF help guide future work in directions that may do so~~will~~JDF.

Chapter 2

Methodology

This section outlines the methodology employed for the reported study. The protocol outlined is referred to as the *Double Reading Procedure* and was first implemented by Fodor et al. (2019). Under this protocol, participants are asked to read aloud visually presented sentences twice, once without taking any time to preview sentence content (Reading 1), and then again after unlimited preview (Reading 2).

Fodor et al. (2019) aimed to investigate the extent to which preview impacted the prosodic phrasing of center embedded sentences, as well as whether or not readers would find the doubly center embedded sentences more comprehensible after preview (or, comprehensible at all, as the doubly center embedded sentences often were not on first attempt). In the prosody literature up to this point, preview has largely been ignored as a factor in reading aloud tasks. Fodor et al. (2019) found that preview did indeed impact both the prosodic grouping that readers used and comprehensibility.

While the questions being in the current study here are different, we are still concerned with the prosody that is produced, as well as the difficulty the reader experiences in parsing a sentence in order to read it aloud. This experimental paradigm eliminates the possible noise of not knowing whether a given pronunciation represents a considered or naive attempt to read a sentence aloud.

2.1 Materials

In total there were 16 experimental items each constructed in 4 versions, and 32 fillers in two versions. The design decisions are discussed in detail in this section.

2.1.1 Experimental items

Table 2.1: Illustrative experimental item, constructed in four versions.

Version	Sentence
D Arg	He had intended to cram the paperwork in the drawer into his boss's desk.
Q Arg	Had he intended to cram the paperwork in the drawer into his boss's desk?
D Mod	He had intended to cram the paperwork in the drawer of his filing cabinet.
Q Mod	Had he intended to cram the paperwork in the drawer of his filing cabinet?

The basic experimental items were created in a 2 x 2 design with one factor being Speech Act (interrogative/Q vs. declarative/D) and the other being PP2 Status, i.e., PP2 was either a PP1 which must be an argument of the verb (Arg) or else one which can be a modifier (Mod) of the preceding phrase (PP1). A full list of experimental items is available in Appendix **??**.

The experimental stimuli were based on earlier pilot study exploring this same phenomenon Peckenpaugh (2016), with several adjustments made to accommodate the objectives of the current study. The sequence of parts for each of the basic items was always the same, shown in (38).

(38)

Introductory material			*Construction*			
Subject	Auxiliary	Matrix verb	Infinitival verb	Object	PP1	PP2
*Order shown for **D** versions; reversed in **Q***					*always ambiguous*	*Disam-biguation (in **Arg** versions)*

All four versions of any given quadruple used the same introductory material, the only difference arising through the necessary inversion of auxiliary and matrix subject, as required by the Speech Act factor. Across quadruples, subjects alternated between *she* and *he*, with half using one and half using the other; the auxiliary was always *had*. The matrix verb did not vary within a quadruple, but did vary between quadruples; for any given quadruple, the matrix verb was one of four verbs of mental state (*decide*, *intend*, *plan*, or *want*).

The verb within the construction did not vary within a quadruple, but a given quadruple could have one of four verbs: *cram*, *put*, *set* or *stick* The construction verb form was always infinitival. Each construction verb appeared in four different quadruples, and was paired once with each matrix verb, to create 16 unique pairings of matrix verb to construction verb. Thus, for matrix verb *decide*, for example, *decided to cram*, *decided to put*, *decided to set*, and *decided to stick*; and for construction verb *cram*, *decided to cram*, *intended to cram*, *wanted to cram*, and *planned to cram*.

The word order and content of the construction was the same across all versions of a quadruple, with the exception of the content of PP2 which varied across the PP2-Status factor: The Arg versions of a quadruple had a PP2 which was headed by *into* or *onto*, while the Mod versions had a PP2 which was headed by *of* or *from*.

PP1 was the same across versions of a given quadruple, e.g., *cram the paperwork in the drawer...* (see Table 2.1's illustrative example). That is, PP1 was identical (and temporarily ambiguous) in every version of a given quadruple, being interpretable as either the goal argument of the construction verb or as a modifier of the object NP. However, in Arg versions of a quadruple, the argument interpretation of PP1 cannot be sustained once PP2 is encountered. In those cases PP2 must fill the goal argument slot and PP1 must be a modifier. The working assumptions about parsing discussed earlier, i.e., that the parser will initially assume PP1 to be the goal argument due to the primary status of arguments, assumes that Arg versions of a quadruple require reanalysis. Between quadruples, the preposition that headed PP1 varied, but was always one which was compatible with it being a goal argument or a modifier of the object: *in* (8), *on* (7), and in one case, *under*.

One benefit of using a complex verb cluster (auxiliary + matrix participle + infinitive) rather than a single verb[1] was that it isolated the differences across the versions of a quadruple triggered by the Speech Act factor to the left extremity of the introductory material of the sentence: only the position of the subject and the auxiliary were affected, meaning that the construction itself was completely untouched by this manipulation.

The purpose of including introductory matrix verbs was to reduce the oddity of the polar interrogative versions of each quadruple. It seems odd to ask, “Did Mary put the jelly beans in the window onto a fancy dish?” because, when it is clear that the speaker already knows so much about the situation, it becomes difficult to imagine a pragmatically plausible context where such a question would be asked. Such sentences might well be described as “prosecutorial[2].” Arguably, this is somewhat mitigated by the addition of a verb like *decided*: rather than asking about facts that we already seem to know, we are instead asking about an actor’s mental state with

[1]Note that the use of an auxiliary also eliminates length differences across D vs. Q versions of a quadruple: if an auxiliary verb were not present, interrogative versions of a basic item would have an extra word, the result of so-called *do*-support, that would not appear in the declaratives (e.g., *he crammed* ... vs. *did he cram* ...?)

[2]Thank you to Dr. Dianne Bradley for making this observation, and for the very clever “prosecutorial” descriptor.

regard to those facts. Even if we know the facts of the situation, we do not necessarily know, for instance, whether it was the result of a decision, some third party's action, or mere happenstance. Another adjustment made in order to make the polar interrogative versions of each quadruple more pragmatically acceptable limited the amount of detail in the experimental sentences, so that fewer adjectives and adverbs were included compared to the items employed in Peckenpaugh (2016), and subjects were always third person nominative pronouns (*he* or *she*).

Importantly, the construction verb was always one which demanded a goal argument. Where some of the verbs used in the items employed by Peckenpaugh (2016) only optionally took a goal, the current study used only verbs which require a goal argument. Verbs that optionally take a goal might result in a parse where PP1 is not immediately incorporated as the goal argument, which would mean that PP2 would not necessarily force reanalysis. Consider the contrasting sub-categorization of the verbs in (39) and (40):

(39) **Optional goal** (*hide*)

The gangsters had hidden the shotguns in a U-Haul truck.

✓ The gangsters had hidden the shotguns.

(40) **Obligatory goal** (*put*)

The gangsters had put the shotguns in a U-Haul truck.

∗ The gangsters had put the shotguns.

A verb like *hide*, as in (39), can take a goal, but can also be used without one. A verb like *put*, on the other hand, as in (40), really must have a goal. The use of verbs that require a goal argument in the current study maximized the likelihood of a robust garden path effect in the Arg versions, when PP2 triggered reanalysis. The four construction verbs used in this study were: *cram*, *put*, *set* and *stick*.

Another important consideration was ensuring that the Arg versions had a PP2 which definitively disambiguated the attachment site of PP1 such that reanalysis was forced. In (41), PP2 is implausible as a modifier of *rocking horse*, but not strictly impossible, and the sentence is grammatical with PP2 modifying it. On the other hand, the use of *onto* in (42) completely disallows the modifier interpretation of PP2 at the syntactic level: a PP headed by *onto* cannot grammatically modify the preceding NP.

(41) She had decided to put the child [$_{PP1}$ on the rocking horse] [$_{PP2}$ on the see-saw].

(42) She had decided to put the child [$_{PP1}$ on the rocking horse] [$_{PP2}$ onto the see-saw].

Where Peckenpaugh (2016) relied on plausibility to force reanalysis, the current study uses syntactic disambiguation, such that the Arg versions always have a PP2 headed by *into* or *onto* which cannot head a PP2 that modifies the NP of PP1. This avoids any noise that might result from discrepancies between individuals' real world knowledge or beliefs. For the Mod items, the head preposition of PP2 was always either *from* or *of*, which are compatible with a parse where PP1 is the goal argument and PP2 is modifying the NP within PP1.

It is worth noting that some linguists (e.g., Den Dikken (2006)) believe *of* is not a preposition in the same sense as *from*, *on*, or *in*, etc., in that it appears to be serving a strictly grammatical or functional purpose, without real lexical content. Importantly, it is also only 2 characters, whereas *into*, *onto*, and *from* (the other possible heads of PP2) are all 4 characters. This is revisited and its possible impact is explored in the results section (section 3.4.3).

To sum up, the experimental items were designed to have limited detail, with either *he* or *she* as the matrix subject. A complex verb cluster, e.g., *had decided to cram* was used to facilitate subject-auxiliary inversion without *do*-support in the interrogatives and limit the difference between items, as well as provide a verb of mental state (*decide*, *intend*, *plan*, or *want*) to support more pragmatically plausible questions. PP1 was always interpretable as either the goal argument or a modifier of the object. PP2 differed across the PP2-Status factor, but not across the Speech Act factor. In the two Arg versions of a quadruple, it was headed by *into* or *onto* and was intended to force reanalysis, under the assumption that PP1 had been incorporated into the parse as an argument, since a PP headed by *into* or *onto* must be interpreted as the goal argument, the position that PP1 would have presumably been occupying in the ongoing parse. For the two Mod versions of a quadruple, PP2 was headed by *from* or *of* and therefore was not expected to require reanalysis, as *from*- and *of*-headed PPs can attach as modifiers of a preceding NP (in this case, the NP within PP1), allowing PP1 to stay in the goal argument slot.

2.1.2 Fillers

There were 32 filler items that ranged in complexity, e.g., some contained embedded finite or non-finite clauses, some contained reduced relative clauses or full relative clauses, and some were simple matrix clauses. Of these 32, 16 were designed to end in a sequence of two PPs, to mirror the experimental items (+PP), while the other half contained no final PPs (-PP). The +PP fillers were unrelated to the -PP fillers. All fillers were designed in two versions: declarative (D) and interrogative (Q). For the +PP fillers, PP1 was an argument in 5 of 16 cases, a modifier of the object in 6 cases, and a modifier of the verb phrase in 5 cases. The distribution of attachment sites for PP2 was the same, except there were 6 that modified the NP embedded in PP1 instead of 6 that modified the object. A full list of fillers is available in Appendix **??**.

Table 2.2: Illustrative filler items, constructed in two versions.

Version	Sentence
D +PP	He had forgotten to try the famous pastry in the restaurant of the fancy hotel.
Q +PP	Had he forgotten to try the famous pastry in the restaurant of the fancy hotel?
D -PP	She had forgotten to report that the clerk was ignoring her request.
Q -PP	Had she forgotten to report that the clerk was ignoring her request?

All filler items had the same sort of introductory material as the experimental items (*he/she* + *had* + past participle verb of mental state). The past participle was either one of the four mental state used for the experimental items (*decide*, *intend*, *plan*, and *want*), or one four additional verbs of mental state: *forgot*, *mean*, *need*, or *remember*, with each of the 8 past participles being used twice in the +PP fillers and twice in the -PP fillers, for a total of 4 times each. This means that a participant would see 6 instances each of *decide*, *intend*, *plan*, and *want*, but only 4 instances of the filler-only mental state verbs. Fillers used both mental state verbs from the experimental items as well as others was to prevent the experimental items as being identifiable by which mental state verb was used, and to avoid extreme amounts of repetition for any given lexical item.

2.1.3 Length

Length was tightly controlled across items. For experimental quadruples, all sentences were between 66 and 75 characters long, and between 13 and 15 words long. The length within a quadruple never varied across the D vs. Q factor. Across the PP2-Status factor, given that the content of PP2 differed within a given quadruple, there was a maximum length difference of one character. Two quadruples varied in word length across PP2-Status by one word. Across all quadruples an equal number were longer (word- and character-wise) in the Arg condition as in the Mod condition. The experimental items ranged from 18 to 22 syllables.

Control over filler pair length was slightly less stringent. They ranged from 63 to 79 characters and 12 to 14 words. Length was never different within a filler pair, since only the Speech Act factor was implemented in the construction of fillers.

2.2 Participants recruitment

All participants in the were undergraduate students enrolled at Queens College in Psychology 101[3] who participated for course credit. Self-reported age ranged from 18 to 25 years. Participants were recruited a software system designed for university participant pools. Students saw a recruitment notice on the system website (see Appendix **??**), and were able to schedule their own appointment time within the hours offered.

The 35 participants recruited were self-identified native and primary speakers of American English. One participant was disqualified post-hoc after producing a Caribbean English pronunciation pattern; one further participant was excluded post-hoc due to an extremely disfluent reading cadence. A final participant was excluded due to a technical issue. All excluded participants were still awarded class credit for participating.

2.3 Location

All data were collected in a private room with only the experimenter and participant present. While every effort was undertaken to ensure a quiet environment, intrusive noise from passersby or neighboring rooms were sometimes unavoidable. This resulted in some unusable or partially unusable recordings (detailed in section 3.4.1

[3]IRB approval number: 2018-0072

of the results chapter).

2.4 Equipment and software

The experiment was presented on a laptop running Windows 10 with stickers on the keyboard labeling relevant keys: the left shift key was labeled *START*, right shift was labeled *NEXT*, and the touch-pad was labeled *DONE*.

The presentation of items and instruction[4] was done using the Open Sesame software (Mathôt, Schreij, & Theeuwes, 2012) which provides a graphical user interface, scripting language, and interpretation of Python code. The system was capable of 10-20 millisecond accuracy, with the display's 60Hz refresh rate being the limiting factor. Key input had a latency of about 10ms.

Recording used a Blue Yeti USB microphone position near the participant's left hand and angled to point at the space in front of the participant's mouth. The angle was adjusted for each participant's height. Audio was recorded at 44.1kHz single-channel quality.

2.5 Versions of the experiment

The experiment was presented in 4 basic versions, with split-half ordering (where the first 24 of the items presented to one group was the second 24 presented to the other) for a total of 8 groups. Each version contained 7 practice items, 3 of which were overt practice and 4 of which were covert practice, as well as one version of each of the 16 experimental and 32 filler items. No version contained more than one version of a given experimental quadruple, or a given filler pair, and each version contained one member of every experimental quadruple and filler pair. Each

[4] Instructions were also provided verbally and via printout, see Appendix **??**.

participant saw the same number of each type of experimental quadruple: 4 D Arg, 4 Q Arg, 4 D Mod and 4 Q Mod. The experimental items were presented in pseudo-random order, interspersed with 1 to 3 fillers. Ignoring fillers, the same version of a different quadruple never occurred in sequence (e.g., after encountering a D Arg, the next experimental item was never another D Arg).

2.6 Procedure

Participants were given a verbal overview of the experimental procedure and then asked to read a one page review of the procedure (see Appendix **??**) before signing a consent form. Participants then sat at the computer and were again walked through instructions before the first practice item was presented.

Participants completed 3 practice items, then consulted with the experimenter before beginning the main portion of the study. The study also contained 4 covert practice items that were not included in any analyses, to allow participants to settle into the procedure before any results were recorded.

Participants used keyboard button presses to navigate the experimental presentation. Each such key-press terminated the current screen, and initiated display of the screen that was programmed to follow. The succession of 4 screens constituting the presentation of any item was participant-paced, as was the progress from item to item. Between items, the display defaulted to a fixation screen showing a line of ten pluses aligned with the left edge of the to-be-revealed sentence. This was designed to direct the participant's attention to the beginning of the sentence, and thus minimize unintended look-ahead (the issue of potential look-ahead is discussed at greater length in section 2.6.1). Items were uniformly presented without line breaks.

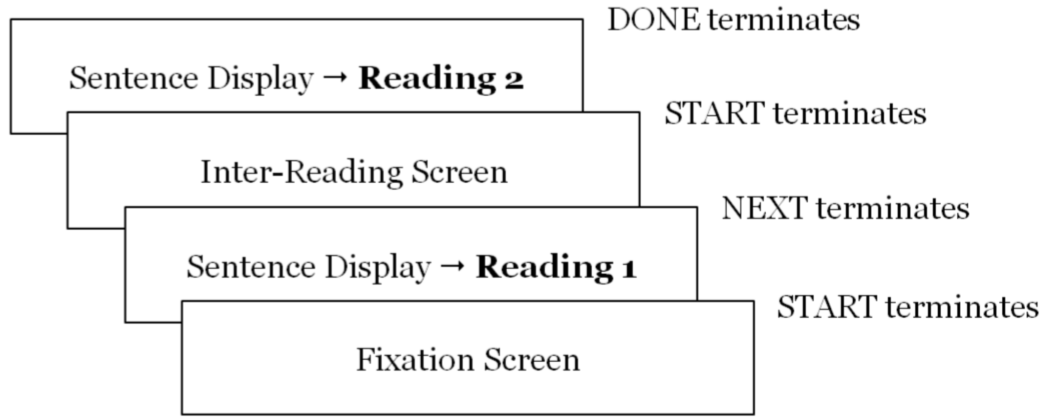

Figure 2.1: Diagram of 4-screen sequence presented for each item, showing the key presses triggering movement between successive screens.

The first *START* key-press that terminated the fixation screen and initiated the first display of a given item also began the first of the 2 recordings collected for each item. That recording continued through the presentation of the inter-item instruction slide and was terminated by the press of *START* that terminated the inter-item instruction slide and initiated the second display of the item.

The second display of an item was displayed in black font on a pale blue background. All other screens were displayed in black font on a pale green background (i.e., fixation screen, first display of an item, and inter-item instruction screen, as well as the initial instructions).

The inter-item instruction slide which was displayed after the first display of an item was terminated contained the following text:

> Your first reading is complete.
>
> Press START to begin the second reading.

The shifting of required key presses and the changing background color were intended to prevent accidental double-presses of any given button from having unintended side effects, and to help the participant track where in the protocol a given screen was located. It took some time for the participants to adapt to the procedure, but generally the necessary habits were acquired before the first item of the experiment proper was presented.

2.6.1 On look-ahead

An advantage that the Double Reading Procedure has is that it allows for certain assumptions to be made about Reading 2 that otherwise would be unclear: Reading 2 certainly represents a *considered* reading of the sentence. Not only has the reader had ample time to examine the sentence, but has necessarily read it and heard it read in producing Reading 1. This means Reading 2 can plausibly be thought to represent a considered prosodic structure, at least more so than an entirely naive reading, and should not reflect any processing issues; a parse should have already been developed during Reading 1, or during subsequent study of the sentence prior to Reading 2.

The nature of Reading 1 is less clear. Because there is variability in the delay between the display of the sentence and the onset of phonation, it is possible that Reading 1 is not entirely delivered without preview. The properties of these Reading 1 delays are discussed at length in a later section, but for now it suffices to say that the very limited preview is possible during a delay that typically falls in the 0.2 to 2.7s range (median = 1s, SD = 0.4). As an example of common reading rates, Ashby, Yang, Evans, & Rayner (2012) reported faster readers as averaging 328 words per minute (wpm), and slower readers 228wpm, in silent reading. That study found that reading time is slower for reading aloud, and that the availability of parafoveal information (i.e., the difference between 1 word and 3 word windows) is less impactful for that reading mode. Given that the experimental items range from 13 to 15 words, most of the R1 delays would not allow even a fast reader to read the entire sentence: the median 1s R1 delay would allow a fast reader time to read very few words; keep in mind that the window is even shorter, because in addition to just reading, the subject is also handling several other cognitive processes (e.g., visual processing, lexical access, issuing motor commands, etc.). The utterance of Reading 1 should, therefore, contain within it any behavioral reflex of whatever parsing difficult the reader has, for most recordings.

In order to clearly understand the results of this double reading study, it is important to understand the mechanics of reading. Specifically, we would want to know at what point during the reading of a temporarily ambiguous sentence the participant will become aware of the existence of a disambiguating PP2, since this is when it will be realized that the initial parse may well crash. The work of several decades on this subject is thoroughly summarized in Rayner, Pollatsek, Ashby, & Clifton (2012). They describe reading as consisting of a series of fixations, when foveal vision takes in a small region of the visual field, and saccades, where the eyes move ahead ballistically (i.e., on a planned trajectory that cannot be interrupted). As a consequence of the ballistic property of saccadic movement and the additional finding that landing sites (fixations) are not random, we can infer that at least some look-ahead is available, i.e., a reader must know something about what is coming in order to plan a suitable landing site. The primary predictor of fixation point seems to be the character length of a word, meaning that the presence of characters and word boundary information (represented orthographically by spaces in languages like English) at least are necessary at the periphery of attention, i.e., within the perceptual span. Some details on the perceptual span, or the information that can be accessed by the eyes at any given time, is discussed in brief, with special attention to its relevance for the study at hand.

Rayner et al. (2012) discuss a number of studies that explore the size and properties of this span, the most fruitful of those studies being based on a gaze-contingent moving-window technique. In this technique, text is presented on a video monitor while the reader is also hooked up to eye-tracking equipment. A computer constantly samples the position of the reader's eyes and updates the display accordingly. Using this elaborate system, and the mutilation of text outside a window of clear text, a so-called moving window around the reader's point of fixation is created. By manipulating the size of this window, it was found that reading speed is maximized when about 15 characters to either side of the fixation site is available (it turns out this is actually asymmetric, and the window need only go as far as the start of currently fixated word in the direction of what has already been read, i.e., to the left for English readers).

In order to determine what information was available at the periphery of the perceptual span, the amount of information outside a window of clear text known to be smaller than the ideal (e.g., 21 characters, 10 to either side) was manipulated. When all characters and spaces were replaced with *X*, essentially destroying all information outside the window, reading was slower than when character spaces were maintained, but all other information was obscured. Improvements in reading speed also occurred when characters were replaced with characters that had similar shape (i.e., the same pattern of ascenders and descenders) as the character they replaced, with and without spaces. Using these techniques and manipulating the size of the window, they were able to determine that it is only word boundary information that is available at the extreme edge of the perceptual span; character shape (ascenders and descenders) is available about 10 characters out from the fixation point, and character identity is available more or less only for the fixated word.

The relevant question for the study at hand is as follows: how much of the sentence will the reader have seen and processed when a given word is being spoken? A typical item is displayed in (43), with the words expected to be fixated underlined, numbered by presumed fixation sequence, and labeled. The number of characters (including spaces) intervening before the start of the disambiguating region (the left edge of PP2) is displayed below each label. These counts are calculated from the initial character of the fixated word to the initial character of the disambiguating region; the actual fixation site is likely to be closer to the center of the word, meaning the distance would be shortened by a few (1-4) characters, depending on the length of the fixated word.

(43)

He had	<u>intended</u>	to	<u>stick</u>	the	<u>letter</u>	in the	<u>mailbox</u>	onto the	<u>proper stack.</u>
	1-INITIAL		2-VERB		3-OBJ		4-PP1	DISAMBIGUATION-PP2	
	45		32		22		7	0	

Table 2.3 describes these distances across items all experimental items. Note that these values do not vary across condition, because counting starts after both the subject and auxiliary verb, and ends before PP2, and the only changes across versions are subject-auxiliary inversion and the content of PP2.

Table 2.3: Distance in characters from fixation to disambiguation in experimental items.

	1-INITIAL	2-VERB	3-OBJ	4-PP1
Median	46	34.5	25.5	7.5
Maximum	50	38.0	27.0	9.0
Minimum	45	32.0	21.0	5.0

From the initial fixation point, the distance to disambiguation ranges from 45 to 50 characters, with a median of 46 characters. If we recall that word boundary information is available 15 to 18 characters to the right of fixation, we can be certain that the disambiguating region is far out of view until several fixations in.

When does the reader become aware of the existence of PP2? When fixated on the direct object head noun, the range of distance is 21 to 27 characters, with a median of 25.5: PP2's content is still outside of view, even in the case of the smallest distance, and adjusting it to be a few characters smaller to account for the fact that fixation is likely to occur closer to the center of a word rather than on its first character. At most, the presence of the first few characters of PP2's preposition may be available, but certainly not the character space after it. The distance from the PP1 fixation point (the head noun within that PP) ranges from 5 to 9 characters, with a median between 7 and 8 characters. Thus, we can say with some certainty that the reader of a sentence such as (43) will be aware that another phrase, one which starts with a 4-character word, remains to be incorporated into the parse sometime after processing of the direct object, and before processing of PP1.

There is yet another piece to consider: the so-called eye-voice span (EVS), and the fact that the readers in this study are reading aloud rather than silently. According to Laubrock & Kliegl (2015), when reading aloud the voice is typically behind the eyes by some 10-20 characters (M = 16.2 characters, SD = 5.2 characters). Adjusting Table 2.3 by subtracting 16 from each cell, we can approximate the position of the voice when the disambiguating region comes within the perceptual span. These values are shown in Table 2.4.

Table 2.4: EVS-adjusted character distance to disambiguation in experimental items.

	1-INITIAL	2-CONSTRUCTION VERB	3-OBJ	4-PP1
Median	30	18.5	9.5	-8.5
Maximum	34	22.0	11.0	-7.0
Minimum	29	16.0	5.0	-11.0

It is likely, then, an oral reader's voice would actually still be on the object when the eyes' fixation begins to provide information of some kind about the existence of PP2, and will still be pronouncing PP1 when the eyes are first fixated on PP2. This raises a question about any prosodic breaks produced after the object, because it is difficult to distinguish between an intentional prosodic break at that point, and one arising from the reader using a natural break for hesitation related to the garden path effect of discovering the disambiguating PP2.

2.7 Measurements of utterance timing

The elicitation protocol described above asked participants to read each sentence twice, once with no preview at all (Reading 1), and then again without any time pressure (Reading 2). Reading 1 (R1) delay is the elapsed time after a sentence is first displayed and when the participant begins speaking. Reading 2 (R2) delay is the same measure, but from the start of the second recording, which begins after the key press that terminates the inter-item instruction slide. Inter-reading time (IRT) is a measure of the time elapsing between when a participant stops speaking after R1 and when speaking resumes for R2. IRT encompasses but is not synonymous with R2 delay, because IRT also includes the elapsed time after the participant stops speaking and the end of the first recording. In this way, IRT is measured across both recordings.

The process for measuring makes use of Voice Activity Detection (VAD) software, which reports whether a given interval in a sound file contains speech-like noise. It's worth making clear that while VAD is employed, most of the measurements of interest are actually the inverse, i.e., the amount of time in a recording that does not contain speech-like noise. For each recording, the amount of time elapsed from the beginning of the recording to phonation onset and offset was found using VAD; then, R1 delay, R2 delay, utterance length and IRT were calculated as a function of each recording's length and the VAD-reported onset and offset of phonation.

The specific software used included a homemade Python script and Google's

WebRTC VAD. The recordings were 44.1kHz WAV files down-sampled to 8kHz via SOX[5]. Google's VAD system used Gaussian Mixture Models to make probabilistic decisions as to whether a given audio frame was speech or noise (see Falk & Chan (2006) for a complete description). Google's implementation takes one parameter called aggressiveness: a 4-tier setting for the level of confidence necessary to call a given interval speech. The implementation codes this setting on a 0-3 scale, where 0 is the most lenient (most likely to label a frame as speech) and 3 is the most stringent (most likely to label a frame as noise).

The recordings vary in the volume of the speaker's voice and the amount of background noise present. An algorithm was constructed to allow for the most stringent (highest VAD aggressiveness) measurement of the least modified data that gave plausible measurements. Specifically, each file was measured using the highest possible aggressiveness for the VAD algorithm and no modification of the recording. If the timings detected were not plausible, the timings were re-measured with the same rejection rate, but after the recording had undergone a 200Hz high-pass filter[6] (HPF). If that still failed, a 400Hz HPF was used. After a further failure, the VAD aggressiveness was lowered, with each HPF value tried again (0, 200Hz, 400Hz); and that process was itself repeated until the lowest possible rejection rate was tried of the four possible settings. The majority of measurements were collected using the highest aggressiveness (85.4%), with more than half requiring no HPF (59.6%) and most of the request requiring a 200Hz HPF (40.1%).

A plausible set of measurements was required to meet the following criteria:

A. *Utterance length:* An utterance length between 2s and 10s, where utterance timing is the longest contiguous span in the recording that VAD reports as

[5]Google's VAD API only accepts WAV files with sample rates that are a multiple of 8kHz. It ultimately down-samples all files to 8kHz, regardless of the input sampling rate.

[6]The exact algorithm is available on github (URL: bit.ly/2uMrcrG)

phonation, with breaks in phonation of less than 1s not breaking contiguity, as Goldman-Eisler (1961) found that a large majority (82.5 to 87%) of pauses in fluent speech are less than 1s. Stimuli range from 18-22 syllables in length. If we assume a speech rate of 3 to 7 syllables per second (Jacewicz, Fox, & Wei, 2010) we would expect utterances between 2.5s and 7.3s. Conservative thresholds higher and lower than the expected were used, especially on the higher end, to allow for any difficulties processing or fluency that might have lead to longer reading times.

B. *Minimum leading silence:* A leading silence ("delay") of more than 120ms. Even a very fast human reaction time should not permit a delay shorter than 120ms, so a shorter delay likely means an inaccurate set of measurements has been reported.

C. *Maximum edge silence:* A maximum trailing and leading silence length of less than 95% of the file's length was also used, in order to filter out recordings that do not represent a valid trial. Very long silences less than this very conservative threshold that impact the IRT are dealt with in the data clean-up rather than via phonation detection, as described in the results section of this paper (section 3.4.1).

With 32 participants reading 48 items (experimental and filler) twice each, there are an expected number of 3072 recordings; due to technical issues at the time of data collection, 71 recordings are missing. Of the 3001 recordings subjected to this treatment, 2976 resulted in plausible timings[^handset]. A review of those that did not result in plausible timings found 9 recordings that were too noisy for computer analysis, but still usable, and those timings were recorded by hand.

To verify the accuracy of the computer measurement, timings were collected by hand for 240 recording. There was a significant positive correlation between hand-measured and computer-measured timings ($r(118)=0.87$, $p < 0.001$), with a median difference of 0.4s[7] (SD = 1.5).

[7] Hand measurement was done to the nearest half second, so a fair amount of error is to be expected.

2.8 Prosodic judgments

A trained linguist informant naive to the research being conducted listened to recordings and reported the presence or absence of breaks in certain regions[8] of the sentence, as well as several other judgments. She was instructed to familiarize herself with a speaker's speech patterns before rating any recordings by listening to 6 filler item recordings from that speaker. She was given a diagram of the sentences as in (44), as well as full plain-text lists of all items.

(44) $$\text{She had } \underset{\text{V}}{\underline{\text{wanted to set}}} \overset{\text{V}}{\%} \underset{\text{OBJ}}{\underline{\text{the textbooks}}} \overset{\text{OBJ}}{\%} \underset{\text{PP1}}{\underline{\text{on the top shelf}}} \overset{\text{PP1}}{\%} \underset{\text{PP2}}{\underline{\text{into the file box.}}}$$

She was asked to report on whether or not she heard a prosodic boundary directly following the region labeled **OBJ**, and directly after the following labeled **PP1**. The following definition of prosodic break was provided:

> Please work with the assumption that "prosodic boundary" in what follows is any subset of the following features, clustered in such a way as to trigger your intuition that a new prosodic element (of any size) is beginning: pitch change, volume change, segmental lengthening, or pause.

The judgments requested also included whether or not the speaker struggled, where that struggle began, whether or not the speaker used question intonation, and which break(s) were stronger or more prominent than which other break(s).

Detailed instructions on the order in which items should be listened to, both within speaker and across speakers, were also provided. The result was that she never listened to both readings of a sentence in sequence; she never listened to 2 Reading 1 versions of different sentences in sequence; and she never listened to the

[8] She also reported on a break after the construction verb, but that break was so rare that it is ignored throughout this report.

sentences in the same order for a given participant as she did for the previous one.

Details on the instructions given and the judgments collected can be found in Appendix **??**.

This strict procedure was implemented to hinder the informant from recognizing any patterns in the data, e.g., a systematic difference between Readings 1 and 2. It also mitigated any ordering effects that might occur in the data or as a result of the informant's own process. The familiarization process via filler items allowed the informant to judge the existence of breaks relative to the typical cadence and fluency of a given speaker, prior to exposure to any of the experimental items for that speaker.

2.8.1 Reliability

A second trained linguist repeated the task over 120 recordings selected from 8 participants (two from each group, one per ordering). Even number experimental items were used from 4 participants, and odd numbered from the other 4. There were 8 recordings missing from the 128 selected, so the reliability task resulted in judgments over 120 recordings. The first informant also blindly re-rated those 120, with the recording name obscured and instructions not to revisit her original ratings. Reliability scores (percent of recordings agreed upon) are reported in Table 2.5.

The lower intra-rater agreement for relative break strength was likely impacted by the method of reporting: because the informant was actually asked to provide judgments over three break locations (the third, V, is omitted throughout this report because it was extremely rare, occurring in just over 8% of recordings). As such, disagreement on that break and the fact that break strength is actually a compilation of two judgments (weakest and strongest break) amplified the noise to some extent.

Table 2.5: Inter and intra-rater agreement.

	OBJ	PP1	Break strength
Inter-rater	65.0% K = 0.17** (z = 2.61)	78.3% K = 0.09 . (z = 1.86)	54.2% K = 0.25*** (z = 3.99)
Intra-rater	77.5% K = 0.52*** (z = 5.73)	85.0% K = 0.52*** (z = 5.82)	72.5% K = 0.44*** (z = 5.70)

Note:

*** p < 0.001; ** p < 0.01; * p < 0.05, . p < 0.1

Chapter 3

Results and discussion

This section reports various descriptions and analyses of the recordings obtained, and the relevance of those findings to the research questions motivating this study. The reported results include the effect of Speech Act (declarative/D vs. interrogative/Q) and PP2 Status (argument/Arg vs. modifier/Mod) on the location of prosodic breaks, as well as on time spent reflecting upon a sentence between readings, which I call inter-reading time (IRT). In order to evaluate the extent to which participants adhered to the protocol as intended, i.e., began to read immediately for Reading 1 as opposed to producing a considered reading in Reading 2, the delay for which a sentence is displayed before a participant begins to read it is compared for Reading 1 (R1 delay) vs. Reading 2 (R2 delay). The prosodic patterns for participants with especially fast and especially slow R1 delays are presented as a way of investigating the extent to which individual differences might impact those patterns, and as a further exploration of the success of the protocol instructions in producing the intended behavior. A finding on the apparent processing cost of interrogative context when compared to declarative context among the filler sentences is also reported.

3.1 Data for analysis

Data for 32 total participants were analyzed. Given 4 versions of the experiment and 2 possible orderings there would ideally be 4 participants per version-order combination. Ultimately, 3 participants had to be excluded for different reasons, resulting in the distribution is as shown in Table 3.1[1]. Participants were removed for the following reasons: one for use of a non-standard dialect, one for extremely disfluent oral reading, and one who was missing more than half of the expected recordings because of a system crash during the procedure.

Table 3.1: Number of participants per version-order combination.

	Order		
	1	2	Sum
Version 1	5	4	9
Version 2	4	4	8
Version 3	4	4	8
Version 4	2	5	7
Sum	15	17	32

Some of the expected 3072 recordings (32 participants x 48 items (16 experimental and 32 filler) x 2 readings) were not used due to intrusive noise during the recording session. Additionally, data were also excluded from analysis if any (Reading 1/Reading 2 pair) was missing; there were 9 such incomplete pairs excluded. Without analyzable data from both members of a pair, it is difficult to determine the extent to which the elicitation protocol was executed as intended (i.e., the extent of preview for Reading 1 vs. Reading 2).

For experimental items, 978 recordings were subjected to prosodic analysis, constituting 95.6% of the utterances elicited. Because IRT data considered utterances in pairs (Reading 1/Reading 2) rather than separately, the database for

[1]The two 5-count cells include 2 additional participants whose data were collected in pursuit of another full set (i.e., towards an expansion to 40 participants) that was not completed due to a lack of participant sign-ups.

response timing took in 489 data points.

Table 3.2: Number of recordings analyzed, as a function of Speech Act and PP2 Status.

	D	Q
Arg	244	240
Mod	246	248

3.2 Prosodic break patterns

The section will report the prosodic phrasings found in the recordings collected, and the extent to which those patterns are or are not influenced by the design parameters of the study (Speech Act and PP2 Status), as well as which reading (Reading 1 or Reading 2) the recording represents. This reported first descriptively (i.e., in terms of frequency), and then using regression models to calculate the statistical significance of whatever effects are found. Finally, a summary of findings and their implications for the hypothesis motivating this study is provided.

In what follows, the distribution of OBJ breaks and PP1 breaks are reported as a function of the four sentence types created by the materials design (D/Q x Arg/Mod), for each of Reading 1 and Reading 2. Then, the patterns of breaks over the two positions are considered, before moving to statistical analysis. Note that while breaks after the construction verb were reported, these breaks were exceptionally rare and occurred in only 8% of recordings, so they have been set aside. The break locations are indicated with a % symbol in (45).

(45)

			V		OBJ		PP1	
	She had	wanted to set	%	the textbooks	%	on the top shelf	%	into the file box.
		V		OBJ		PP1		PP2

As noted in section 2.8, the results reported are based on the subjective judgments of a trained linguist who was naive to the purposes and hypotheses underlying the research.

3.2.1 Individual break patterns

Table 3.3: Percent occurrence of OBJ break (frequency of occurrence in parenthesis) as a function of sentence type and Reading.

	Reading 1		Reading 2	
	D	Q	D	Q
Arg	57.4% (70)	56.7% (68)	73.0% (89)	74.2% (89)
Mod	77.2% (95)	76.6% (95)	84.6% (104)	72.6% (90)

The presence of the OBJ break was sensitive to both Speech Act and reading, with Reading 2 showing a different distribution across the D vs. Q distinction than the Reading 1 recordings.

Table 3.4: Percent occurrence of PP1 break (frequency of occurrence in parenthesis) as a function of sentence type and Reading.

	Reading 1		Reading 2	
	D	Q	D	Q
Arg	121 (99.2%)	119 (99.2%)	121 (99.2%)	117 (97.5%)
Mod	84 (68.3%)	85 (68.5%)	84 (68.3%)	83 (66.9%)

The PP1 break was almost always present for cases where PP2 was an argument; and it was present substantially less often, but still there a majority of the time, for cases where PP2 could be interpreted as a modifier. Speech act and reading did not appear to impact the overall distribution of the PP1 break.

3.2.2 Combined break patterns

When looking at both breaks together, a sentence could have one of four patterns: both the OBJ and PP1 break present; only OBJ present; only PP1 present; or neither break present. There were only 5 cases where neither was present, and those were omitted in the tables of prosodic patterns.

Table 3.5 shows that for Arg sentences, there are very few instances with the

Table 3.5: Percent occurrence of both breaks as a function of sentence type and Reading.

	Reading 1				Reading 2			
	Mod		Arg		Mod		Arg	
	D	Q	D	Q	D	Q	D	Q
OBJ only	31.1%	31.4%	0.8%	2.5%	31.7%	30.9%	0.8%	0.8%
Both	54.1%	43.0%	72.1%	71.7%	45.5%	46.3%	56.6%	55.8%
PP1 only	14.8%	25.6%	27.0%	25.8%	22.8%	22.8%	42.6%	43.3%

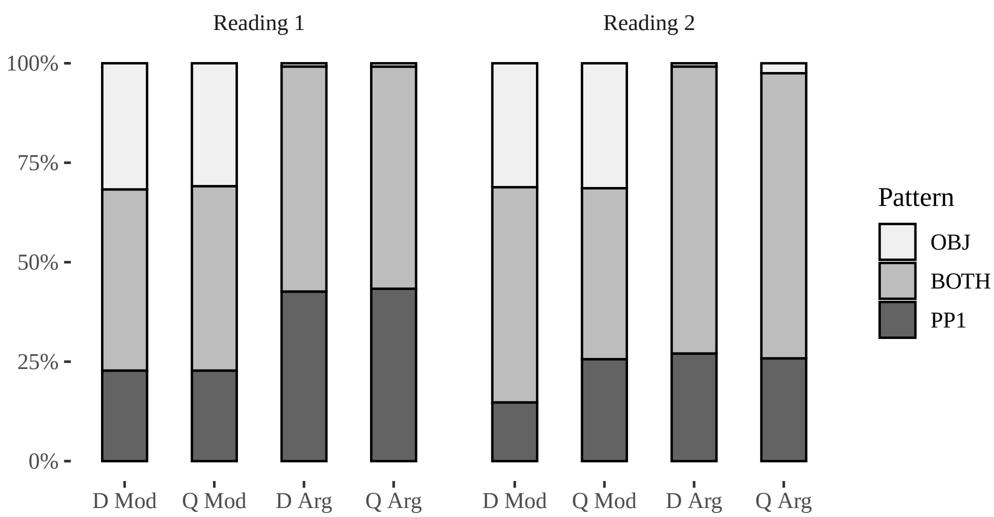

Figure 3.1: Break pattern as a function of sentence type and Reading.

OBJ-only pattern (0.8% in declaratives, 2.5% in interrogatives); whereas that pattern is fairly frequent for Mod sentences (31.1% in declaratives, 31.4% in interrogatives). The pattern with both breaks is somewhat more common for Arg sentences (72.1% in declaratives, 71.7% in interrogatives) than Mod (54.1% in declaratives, 43.0% in interrogatives). The PP1-only pattern occurs at about the same rate in Arg interrogatives (25.8%) as in Mod interrogatives (25.6%) and Arg declaratives (27%), but is noticeably less common for Mod declaratives (14.8%). These proportions are visually represented in figure 3.1.

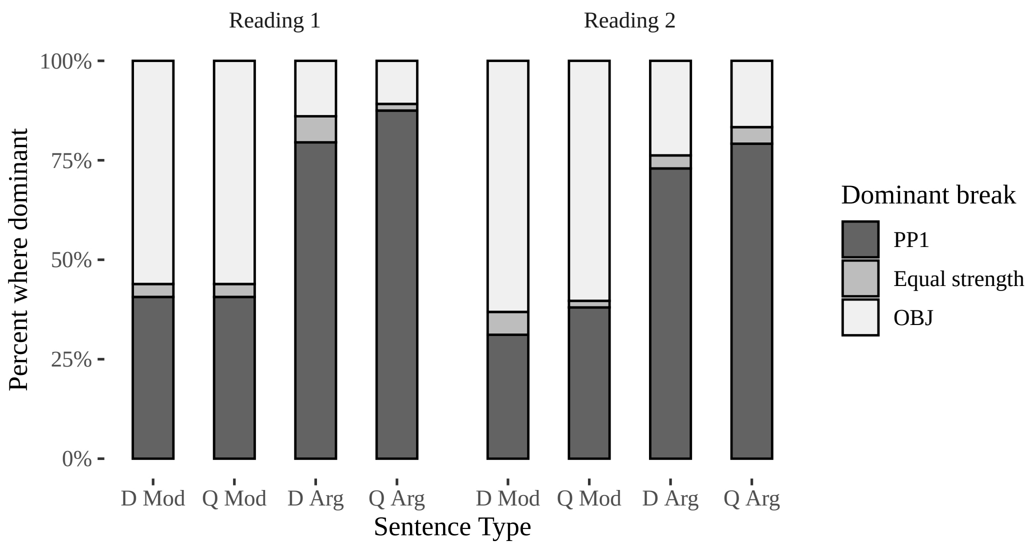

Figure 3.2: Percent break dominance occurence as a function of sentence type and Reading.

3.2.3 Break Dominance

The relative strength of the PP1 and OBJ breaks was also collected. Figure 3.2 incorporates this information, where "PP1 dominance" means that the PP1 break was reported to be stronger than the OBJ break; "OBJ" dominance means the opposite; and "Equal strength" means that neither break was reported to be stronger than the other (the 5 instances with no breaks were again omitted).

When looking at the combined break patterns, one can think of there being three bins: the PP1 bin, the OBJ bin, and a neutral bin between them. In Section 3.2.2, the neutral bin containing instances of both breaks occurring is robust. This break dominance analysis distributes most of those cases that have both breaks into either the OBJ or PP1 bin, depending on which break is more prominent. When the breaks are of equal strength, they remain in the middle bin, but there are much fewer such cases when looking at dominance instead of simple occurrence.

Figure 3.2 clearly shows a robust effect of PP2 Status on break dominance, and little to no impact of Reading or Speech Act.

3.2.4 Regression models of prosodic break patterns

A number of mixed effects logistic regression models support the general observations above. Models predicting PP1 break, OBJ break, PP1 break dominance and OBJ break dominance are reported. All models include crossed random effect intercepts (participant and item), but due to convergence errors, no random slopes for any predictors are included.

The intercept always represents the Mod sentence type, which is not expected to present any particular difficulty to the reader, since the Mod PP2 Status is compatible with what is assumed to be the running parse when it is encountered (i.e., PP1 has been interpreted as the goal argument of the verb, and PP2 does not disrupt that interpretation). For those models where Speech Act is included in the model, the intercept represents the declarative sentence type. In this way, the more complex sentence types are compared to the simplest available in the model. If Reading is included in a model, the intercept represents Reading 1.

For each analysis, a reduced model and the full model (i.e., the model containing all predictors of interest) is reported. In each case, the reported reduced model is the one with the lowest reported Akaike Information Criterion[2] (AIC) from the set of models that include any subset of the following predictors: Speech Act, PP2 Status, Reading, and the interactions between Speech Act, PP2 Status and Reading. This method of model selection is consistent with the proposal of Wax & Kailath (1985).

Model comparisons did not always find significant differences between the more complex models, but in each case, the selected model was compared to a minimal

[2]AIC is a representation of the amount of information lost by using a regression model to estimate data points. It is a measure that balances both the goodness of fit of a model and the simplicity of a model, guarding against over fitting and under fitting the data involved.

model where where fixed effect variables were removed, leaving only an intercept, and all reported models represent improvement over the minimal model to a statistically significant degree. That comparison is reported for each model. All regression models were run using the lme4 R package (Bates, Maechler, Bolker, & Walker (2019)), with p-values calculated via the lmerTest R package (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen (2019)).

3.2.4.1 Break occurrence

In the full model predicting OBJ break occurrence, shown in Table 3.6, only the estimate for D Mod Reading 1 (the intercept) and the effect of PP2 Status show statistical significance.

Table 3.6: Mixed effects logistic regression model predicting OBJ break occurrence (FULL).

Outcome: OBJ break (FULL)	Estimate	Std. Error	p
D Mod, Reading 1 (Intercept)	0.70	0.09	< 0.001
Q	0.11	0.11	0.34
Arg	-0.28	0.11	< 0.05
Reading 2	0.07	0.05	0.15
Q:Arg	-0.14	0.16	0.39
Q:Reading2	-0.11	0.07	0.12
Arg:Reading2	0.08	0.07	0.27
Q:Arg:Reading2	0.13	0.10	0.19

Table 3.7 shows a reduced model, predicting the occurrence of an OBJ break with estimates for the coefficients of the fixed effects of Reading 2, PP2 and the interaction between Reading and PP2 Status. The removal of other predictors allowed the Reading x PP2 Status to become a significant predictor. A comparison between the reported model and a minimal one found that the reported model was better with a high level of confidence (AIC_{MIN}=1068.0, AIC_{BEST}=1031.6, $\chi^2(2)$=30.5, $p < 0.001$).

Table 3.7: Mixed effects logistic regression model predicting OBJ break occurrence (REDUCED).

Outcome: OBJ break (REDUCED)	Estimate	Std. Error	p
D Mod, Reading 1 (Intercept)	1.39	0.45	< 0.01
Reading 2	0.11	0.23	0.62
Arg	-1.98	0.50	< 0.001
Reading 2 x Arg	0.81	0.32	< 0.05

The log odds[3] of an OBJ break for Mod Reading 1 is 1.39 (std. error = 0.45, $p < 0.01$). The log odds of that break increased in Reading 2 but the increase was not statistically significant. PP2 arguments reduced the log odds of an OBJ break compared to PP2 modifiers by a robust amount, but less so in Reading 2 than in Reading 1.

The OBJ break is expected to occur more often in Mod cases, because that break marks the argument attachment (and therefore a change in branching direction) of PP1.

The full model for predicting PP1 also showed significance only for the intercept and the effect of PP2 Status.

Table 3.8: Mixed effects logistic regression model predicting PP1 break occurrence (FULL).

Outcome: PP1 break (FULL)	Estimate	Std. Error	p
D Mod, Reading 1 (Intercept)	0.68	0.07	< 0.001
Q	0.01	0.09	0.89
Arg	0.31	0.09	< 0.001
Reading 2	0.00	0.04	1.00
Q:Arg	0.00	0.13	0.97
Q:Reading2	-0.01	0.06	0.82
Arg:Reading2	0.00	0.06	1.00
Q:Arg:Reading2	0.00	0.08	0.97

[3]Log odds is, in this case, the natural log of the odds ratio, so the log odds of A is $\log_e(P(A)/P(\neg A))$. A log odds of 1.39 translates to an odds ratio of 4.01:1 (1.39^e=4.01) and a probability of 80% (4.01/(1+4.01)=0.80).

The best model for predicting the occurrence for the PP1 break was one where only PP2 Status was included as a predictor. The chosen model was again significantly better than the minimal model (AIC_{MIN}=855.6, AIC_{BEST}=629.6, $\chi^2(1)=228.0$, $p < 0.001$).

Table 3.9: Mixed effects logistic regression model predicting PP1 break occurrence (REDUCED).

Outcome: PP1 break (REDUCED)	Estimate	Std. Error	p
Mod (Intercept)	0.96	0.30	< 0.01
Arg	4.12	0.44	< 0.001

Sentences with argument PP2s had greatly increased log odds of a PP1 break compared to ones with modifier PP2s. This is again expected, because the PP1 break is indicating the change in branching direction for argument attachment of PP2. That Speech Act is not a relevant predictor is evidence against a prosodic explanation of the motivating intuition for this study; we would expect both a main effect of Speech Act and definitely an interaction between Speech Act and PP2 Status, if the prosody were more or less different across the PP2 Status factor for interrogatives than for declaratives.

3.2.4.2 Break dominance

Models were also run for predicting break dominance. The full model predicting OBJ break dominance is shown in Table 3.10.

Table 3.11 reports the best model for predicting OBJ break dominance. The best model was one with fixed effects for reading and PP2 Status. There was no statistically significant effect of Speech Act on OBJ break dominance.

PP1 break dominance and OBJ break dominance are not entirely complementary, because it is possible for both breaks to have equal prominence. As such, models predicting PP1 were also explored.

Table 3.10: Mixed effects logistic regression model predicting OBJ break dominance (FULL).

Outcome: OBJ dominance (FULL)	Estimate	Std. Error	p
D Mod, Reading 1 (Intercept)	0.50	0.09	< 0.001
Q	0.03	0.12	0.82
Arg	-0.45	0.12	< 0.001
Reading 2	0.07	0.05	0.22
Q:Arg	-0.03	0.17	0.86
Q:Reading2	-0.03	0.07	0.68
Arg:Reading2	0.03	0.07	0.71
Q:Arg:Reading2	0.00	0.11	0.97

Table 3.11: Mixed effects logistic regression model predicting OBJ break dominance (REDUCED).

Outcome: OBJ dominance (REDUCED)	Estimate	Std. Error	p
Mod, Reading 1 (Intercept)	-0.16	0.32	0.62
Reading 2	0.40	0.16	< 0.05
Arg	-2.32	0.18	< 0.001

The full model predicting PP1 break dominance failed to converge, so only the reduced model is reported.

Table 3.12 reports the best model for predicting PP1 break dominance. Unlike the model for predicting OBJ break dominance, the best model for predicting PP1 break dominance includes Speech Act as a predictor. The best model is one with fixed effects for reading, Speech Act, and PP2 Status.

Table 3.12: Mixed effects logistic regression model predicting PP1 break dominance (REDUCED).

Outcome: PP1 dominance (REDUCED)	Estimate	Std. Error	p
D Mod, Reading 1 (Intercept)	-0.19	0.33	0.57
Reading 2	-0.38	0.15	< 0.05
Q	0.31	0.15	< 0.05
Arg	2.20	0.17	< 0.001

This model was better than a minimal model (AIC_{MIN}=1290.4, AIC_{BEST}=1078.8,

$\chi^2(3)$=217.59, $p < 0.001$). PP1 break dominance was much more likely for sentence with argument PP2s than sentences with modifier PP2s, with interrogatives having slightly increased log odds of PP1 break dominance. Log odds of PP1 break dominance were slightly less in Reading 2 than Reading 1. There were no significant interaction terms.

Because reading was a significant predictor for 3 of the 4 models reported, and there are theoretical reasons to believe that Reading 2 is more representative of the natural or intended prosody of the reader, models were also run predicting PP1 dominance and OBJ dominance for Reading 2 data only. In both cases, the best model had the same structure: fixed effects of Speech Act and PP2 Status, with no interaction term.

Table 3.13: Mixed effects logistic regression models predicting break dominance in Reading 2 (REDUCED).

	Outcome: OBJ Dominance			Outcome: PP1 Dominance		
(Reading 2 only)	Estimate	Std. Err	p	Estimate	Std. Err	p
D Mod (Intercept)	0.66	0.24	< 0.01	-0.97	0.27	< 0.001
Q	-0.30	0.22	0.16	0.35	0.22	0.1
Arg	-2.07	0.24	< 0.001	2.15	0.24	< 0.001

For both OBJ dominance and PP2 dominance, the main effect of Speech Act is non-significant, but its inclusion marginally improves the fit of each model. Even when limited to only Reading 2 data, Speech Act does not interact with PP2 Status, which is again supportive of a non-prosodic explanation for the motivating intuition. That there is a robust effect of PP2 Status is reassuring evidence that prosody is sensitive to syntax, and that the study's item construction is motivating the intended parse.

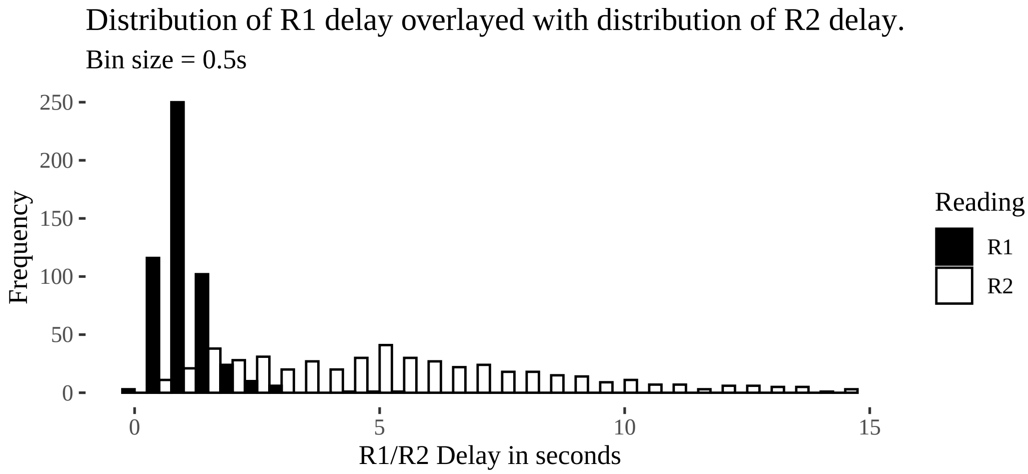

Figure 3.3: Distributions of R1 delay and R2 delay

3.2.5 On Reading 1 delay

Reading 1 (R1) delay is the amount of time between the initial display of a sentence and the start of phonation. Participants' median R1 delay ranged from 0.6s to 1.6s with a standard deviation of 0.25s. The distribution of R1 delay was notably different than that of R2 delay, shown in Figure 3.3 which indicates that participants were adhering to the protocol at least most of the time.

As a way of analyzing the protocol, and the extent to which participants performed as expected, participants were categorized based on their median R1 delay. In what follows, a fast median R1 delay was shorter than or equal to 0.9s, and a slow one was longer than 1.05s, resulting in 12 participants per category. Ten participants had R1 delays between those values, categorized as "normal," and set aside. The calculations for categorizing participants were done over Reading 1 of experimental items (n = 489). Note that while R1 delay category (i.e., fast or slow) is a property of R1 delay, data for both readings is nonetheless explored within these categories.

There is a statically significant difference between the number of cases where both breaks were produced across the fast (44) vs. slow (65) category for Reading 1 ($\chi^2(1)$

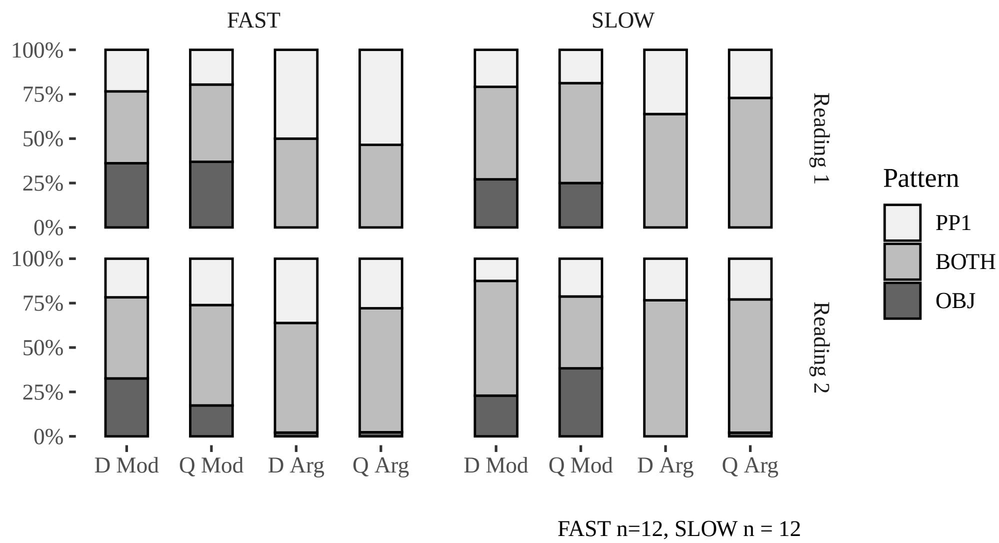

Figure 3.4: Plot of pattern proportions as a function of sentence type.

= 4.05, $p < 0.05$), but not for Reading 2 ($\chi^2(1) = 1.86$, $p = 0.17$). There was also a statically significant difference in the occurrence of both breaks across PP2 Status for Reading 2 in the slow R1 delay category ($\chi^2(1) = 3.97$, $p < 0.05$), but comparisons across other factors represent in Table 3.15 did not yield significant results. The maximum count per cell in the table is 96 (12 participants per category, 8 items per PP2 Status), ignoring missing items.

In cases where R1 delay was small, readers were more likely to produce only one break than if R1 delay was fast. A possible explanation is that the slow starters were more prone to hesitation in general, or perhaps were using both the delay time and their extra break time to look ahead. The significant effect of PP2 Status for slow starters in Reading 2 could be due to the slow starters not fully understanding the Arg sentences prior to Reading 2, thus increasing their likelihood of hesitation.

Table 3.14: Break pattern by sentence type and R1 delay category.

		FAST (n=12)				SLOW (n=12)			
		D Arg	D Mod	Q Arg	Q Mod	D Arg	D Mod	Q Arg	Q Mod
Reading 1									
BOTH		46.5%	50.0%	43.5%	40.4%	72.9%	63.8%	56.2%	52.1%
		20	24	20	19	35	30	27	25
OBJ		0.0%	0.0%	37.0%	36.2%	0.0%	0.0%	25.0%	27.1%
		0	0	17	17	0	0	12	13
PP1		53.5%	50.0%	19.6%	23.4%	27.1%	36.2%	18.8%	20.8%
		23	24	9	11	13	17	9	10
Reading 2									
BOTH		69.8%	61.7%	56.5%	45.7%	75.0%	76.6%	40.4%	64.6%
		30	29	26	21	36	36	19	31
OBJ		2.3%	2.1%	17.4%	32.6%	2.1%	0.0%	38.3%	22.9%
		1	1	8	15	1	0	18	11
PP1		27.9%	36.2%	26.1%	21.7%	22.9%	23.4%	21.3%	12.5%
		12	17	12	10	11	11	10	6

Table 3.15: Occurrence of both breaks as a function of Reading, PP2 Status, and R1 Delay Category.

	Reading 1	Reading 2
Mod		
FAST	39	47
SLOW	52	50
Arg		
FAST	44	59
SLOW	65	72

3.3 Discussion of prosodic break patterns

Throughout the analysis of break patterns, PP2 Status was the most robust predictor of OBJ and PP1 break occurrence and their relative strengths. The OBJ break was more frequent and more frequently dominant for sentences with a PP2 that was an argument than those with a PP2 that could be a modifier; inversely, the PP1 break was more frequent and more frequently dominant for sentences with a PP2 that was interpretable as a modifier than those with a PP2 that was an argument.

Essentially, the expected patterns can be described as in (46) and (47), where % represents a robust prosodic break, and @ represents a weaker break or no break.

(46)

$$\text{... stick } \underset{\textit{dir. object}}{\underline{\text{the letter}}} \ \overset{\text{OBJ}}{@} \ \underset{\textit{PP1}}{\underline{\text{in the mailbox}}} \ \overset{\text{PP1}}{\%} \ \underset{\textit{PP2}_{\textit{Arg}}}{\underline{\text{onto the proper stack.}}}$$

(47)

$$\text{... stick } \underset{\textit{dir. object}}{\underline{\text{the letter}}} \ \overset{\text{OBJ}}{\%} \ \underset{\textit{PP1}}{\underline{\text{in the mailbox}}} \ \overset{\text{PP1}}{@} \ \underset{\textit{PP2}_{\textit{Mod}}}{\underline{\text{of the vice president}}}$$

This is supportive of *hypothesis 1* presented in the introduction; and, in fact, of a broader formulation of it. Essentially, it's predicted that the PP1 break will more often be dominant for Arg cases, while the OBJ break will more often be dominant for Mod cases, because those break locations correspond to the position where the branching direction in the syntactic tree changes.

Hypothesis 1

Argument attachment of PP2 is marked by a dominant prosodic break between PP1 and PP2.

Broadened hypothesis 1

A change in branching direction is marked by a prosodic break.

A change in branching direction is taken to mean the closure of the preceding phrase and attachment into its parent node; in this case, either the closure of PP1 and the attachment off PP2 into the VP or the closure of the object NP and attachment of PP1 into the VP.

That Reading 2 is a significant predictor in 3 of the 4 analyses where its inclusion is possible supports, at least provisionally, hypothesis 2 and 3.

Hypothesis 2

A first reading of a sentence where PP2 is a goal argument will exhibit less natural prosody (more hesitation at and within the PP2 region) than: * A first reading of a sentence where PP2 is a modifier * A second reading of a sentence where PP2 is a goal argument

It is difficult to know whether a given reading represents more or less natural prosody, but given that there is a difference between readings, it seems most likely that Reading 2 is the more natural of the two since it represents a considered reading, rather than a hurried one. *Hypothesis 2-3* are supported only if that assumption is accepted.

Hypothesis 3

A first reading of a sentence with an argument-PP2 will more often be produced with prosodic structure that represents an implausible or ungrammatical parse of the string (PP2 incorrectly attached as a modifier), whereas a second reading sentence will more often be pronounced with the prosodic structure that represents the intended parse (argument attachment of PP2).

(48) *Hypothesis 4*

Reading 1 of a declarative sentence with an argument-PP2 will exhibit less natural prosody (more hesitation at and after the disambiguating region) and be more likely to be produced with prosodic structure that represents an implausible or ungrammatical parse of the string than a Reading 1 of an interrogative sentence with an argument-PP2.

It is surprising that the effect of PP2 Status is generally lessened in Reading 2 when compared to Reading 1, but this can likely be explained as an epi-phenomenon. There is no way to distinguish between prosodic breaks that are intentional and syntactically motivated as compared to those that represent hesitation, a need for a breath, or other factors. It is likely that some of the effect of PP2 Status is actually an increase in hesitation after PP1, and therefore more or longer pauses at that position, which is mitigated in Reading 2. If some readers are, in general, simply producing a break after every phrase, but happen to produce what is perceived as a dominant break after PP1 for the Arg sentences when they are confused, that effect of PP2 Status will go away in Reading 2 once they have had time to figure the sentence out. This might mean that the noise caused by readers that are simply breaking phrase-by-phrase is actually amplified in Reading 2.

That a prosodic break also frequently occurs between phrases when there is no change in branching direction is mitigated somewhat by the fact that such breaks are usually weaker than the ones that do represent such a change. It is likely that these breaks are actually there for non-syntactic reasons; the end of a phrase represents a reasonable time for the speaker to take a breath or pause briefly for processing reasons. It is also likely that some readers are simply producing a break after each phrase.

Speech act is a significant predictor of PP1 break dominance (β=0.31, std. error = 0.15, $p < 0.05$)., but not of any of the other outcomes. It's plausible that the PP1 break is more likely to be dominant in questions than in declaratives because of the need to begin the sentence final rise of question intonation. That there is never an interaction between Speech Act and PP2 Status is discouraging for the hypothesis that it's the *prosody* of questions that make the Arg cases seem easier in the interrogative cases compared to the declarative.

3.4 Inter-reading time

Inter-reading time is the amount of time after the completion of Reading 1 and before the beginning of phonation of Reading 2. The details of how this was measured and defined can be found in section 2.7. IRT is meant to be a measure to some extent of how much difficulty the reader has in processing a given sentence. If a reader spends more time studying a sentence prior to reading it aloud a second time, the IRT will be longer, and I take that has an indicator of processing load.

Importantly, IRT is a measure across pairs of recordings (Reading 1/Reading 2), so the number of data analyzed in this section are half as many as those in the prosody analyses.

3.4.1 Data cleanup

IRTs below 0.25s (2) and above 25.0s (5) were assumed to be implausible and omitted. Experimental data were then Winsorized by participant to bring data below the 2.5% and above the 97.5% threshold to the value at those thresholds. The resulting measure is referred to as wIRT and is distributed as shown in figure 3.5 (n = 489).

Overall mean for wIRT was 6.5s (sd = 3.8). The longest wIRT was 22.2s and the shortest was 0.7s. Median wIRT was 6.1s.

3.4.2 Analysis of IRT data

Figure 3.6 shows the mean IRT as a function of sentence type.

The two slopes are only very slightly divergent. Notably, both Speech Act and PP2 Status appear to have main effects on wIRT, with interrogatives attracting longer (6.7s) IRTs than declaratives (6.4s), and sentences with argument PP2s having

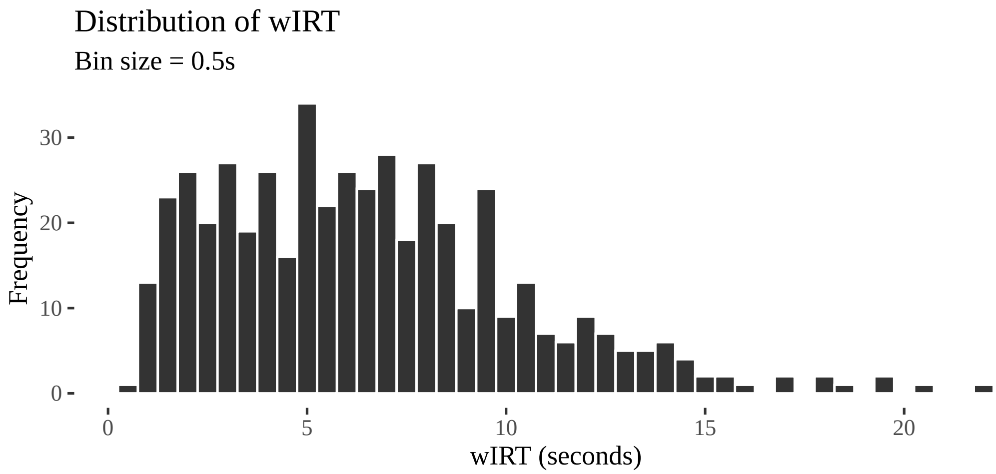

Figure 3.5: Distribution of wIRT.

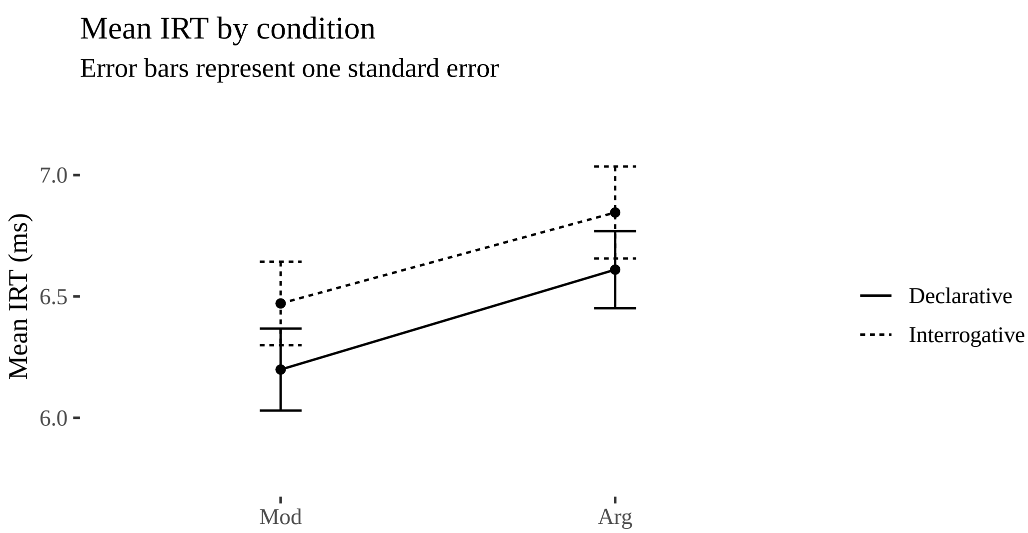

Figure 3.6: Mean IRT as a function of sentence type.

substantially longer IRTs (6.7s) than those with modifier PP2s (6.3s).

Regression models support the observations above. All models discussed include random intercepts for participant and item. Models with random slopes for fixed effects all resulted in singular fits and so random slopes were not included.

By hypothesis, the best model for predicting IRT should be one with fixed effects of Speech Act and PP2 stats and the interaction between them. This model is shown in table 3.16.

Table 3.16: Linear mixed effects regression model predicting wIRT by sentence type with interaction term (FULL).

FULL MODEL	Estimate	Std. Error	p
D Mod (Intercept)	6.32	0.59	< 0.001
Q	0.29	0.30	0.34
Arg	0.42	0.30	0.17
Q x Arg	0.08	0.43	0.85

Of the models with subset(s) of these predictors, the best (the model with the lowest AIC) was the one without the interaction term, shown in table 4.14.

Table 3.17: Linear mixed effects regression model predicting wIRT by sentence type (REDUCED).

REDUCED MODEL	Estimate	Std. Error	p
D Mod (Intercept)	6.30	0.58	< 0.001
Q	0.33	0.21	0.12
Arg	0.46	0.21	< 0.05

The difference between the reduced and full model was not statistically significant (AIC_{FULL}=9103.5, $AIC_{REDUCED}$=9101.6, $\chi^2(1)=0.04$, $p > 0.8$).

3.4.3 On PP2 heads

As mentioned in the items description (section 3.4), half of the items used *of* for the head of PP2 in the Mod cases, while half used *from*. In the Arg condition, half used *into* and half used *onto* to head PP2. An analysis that looks at the identity of the PP2 head found that while *into* and *onto* do not behave differently, *of* and *from* do. Starting from a maximally complex model that included the lexical identity of the matrix verb, the construction verb, the head of PP1, and the head of PP2, as well as Speech Act and PP2 Status, the model that best predicted wIRT was one that is essentially the same as the reduced model just reported (i.e., with Speech Act and PP2 Status as predictors), except it substitutes the lexical identity of the PP2 head for PP2 Status, with *into* and *onto* collapsed into one level used as the reference level (intercept).

Table 3.18: Linear mixed effects regression model predicting wIRT by Speech Act and PP2 head.

PP2 head model	Estimate	Std. Error	p
D into/onto (Intercept)	6.76	0.58	< 0.001
Q	0.32	0.21	0.13
from	-0.08	0.27	0.77
of	-0.84	0.27	< 0.01

Sentences where PP2 was headed by *from* typically had wIRTs that were 0.84s faster than sentences where PP2 was headed by *into*/*onto*; when the PP2 head was *from* wIRT was only 0.08s faster than for *into*/*onto*, a difference that is not statistically significant ($p > 0.7$).

3.5 Discussion of IRT results

It is clear from the above that argument PP2s (ones headed by *into*/*onto*) result in longer wIRT measures. It also appears that interrogativity increases wIRT to a lesser extent, regardless of the PP2 Status. Because the interaction between the two factors is not a significant predictor of wIRT, we are left to assume that wIRT does not represent a behavioral reflex of the intuition that interrogativity makes difficult to process PP2-attachment ambiguities easier.

The difference between *from* and *of* PP2s is potentially a source of noise. The *from* sentences are less clearly disambiguated than the *of* sentences. Where (49) has only one reading, (50) has another possible reading, albeit somewhat implausible: i.e., we could imagine that in (49) *from her brother-in-law* modifies *the cookies*, while in (4) *of the minivan* cannot modify *the bicycle*.

(49) She had intended to put the bicycle on the roof rack of the minivan.

(50) She had decided to cram the cookies in the basket from her brother-in-law.

This lingering ambiguity could have increased wIRT somewhat because the reader, given unlimited time, spent some of that time noticing and then eliminating that possible reading. The difference here can be explained by once again making an appeal to structural parsing vs. structural association: if we imagine that a *from* PP is associated rather than parsed, the reader is free to consider other possible interpretations. Because *of* can be seen as less a preposition and more a functional head, it stands to reason that it would be treated different, as something that must be parsed immediately, i.e., that *from her brother-in-law* is modifying, where *of the minivan* is in some sense more argument-like. This is similar to an established observation about what constituents pro-forms can stand in for: (51) is ungrammatical, but (52) is fine; the same holds for (53) where *to Fred* is an argument, compared to (54) where *for Fred* is a modifier.

(51) ∗ I saw the student of physics and the one of chemistry arguing with each other.

(52) ✓ I saw the student from Texas and the one from Maine arguing with each other.

(53) ∗ Mary gave a book to Fred and did so to John, too.

(54) ✓ Mary signed a book for Fred and did so for John, too.

It could also be a simpler explanation: the fact that *of* is only two characters, whereas *from* and *into*/*onto* are all four characters, participants may have recognized a pattern, wherein they could be sure that a two-character PP2 head meant the sentence did not have the difficult properties of some of the ones with four-character PP2 heads (most notably the *into*/*onto* cases), and eliminated some of the needed study time.

3.5.1 The processing cost of interrogativity

It is worth taking note of the fact that the mean wIRT for interrogative versions (6.7s) of the experimental sentences in the reported study was longer than for the declaratives (6.4s). While this finding was not statistically significant, Peckenpaugh (2016) found that whole-sentence silent reading times for interrogatives were longer than for declaratives; and Mehler (1963) provided a very early report of the processing cost of interrogativity: a so-called kernel sentence, i.e., a simple declarative, was easier to recall verbatim than was a number of sentences that he considered to be syntactic transformations of that kernel sentence (K): negative (N), polar question (Q), passive (P), and combinations thereof: NQ, NP, QP and NPQ. Mehler found that accurate recall was more frequent for K sentences (300/460, 65.2%) than for the other sentences types, with interrogatives (210/460, 45.7%) being recalled accurately at a lower rate than the two other individual transformations (234/460, 50.9% for N; 243/460, 52.8% for P).

The filler sentences in this study were designed in two versions, interrogative (Q) and declarative (D), so as to provide a diagnostic of the interrogative effect on IRT independent of the experimental question. A linear mixed effects regression model predicting wIRT for filler items by interrogativity with crossed random intercepts (participant and item) found that wIRT is increased by 0.4s for interrogatives (std. error = 0.2; $p < 0.05$); declaratives had a mean wIRT of 6.2s, while interrogatives had a mean wIRT of 6.6s. Half of the fillers had a sequence of two PPs at the end to mirror the experimental items: a model predicting wIRT by the presence of those PPs found minimal effect on wIRT (β=0.01, std. error = 0.22, $p = 0.96$).

Interrogative status itself appears to increase the time needed for participants to feel they have satisfactorily studied a sentence in order to read it aloud correctly. This is consistent with the Mehler (1963) and Peckenpaugh (2016) findings that interrogatives are in some way more complicated or difficult than declaratives.

Chapter 4

General discussion

This chapter will review the questions motivating this study and discuss the extent to which those questions are answered, or not. It will then go on to develop further questions, and propose further studies to explore those new questions and the ones left unanswered here. Finally, it will summarize the findings and the current standing of this area of research.

Recall that the primary motivation for this study was the possibility that PP-attachment garden paths are easier to understand for speakers of American English when presented in the interrogative, as opposed to the declarative. In Section 1.1, a distinction was made between the intuition first outlined in Peckenpaugh (2016) and the current hypothesis. For terminological clarity, recall the definitions originally given in 1.1:

(55) *The 2016 intuition:* Certain pragmatically disambiguated prepositional phrase (PP) attachment ambiguities which are difficult to parse in the declarative are less difficult to parse when presented as yes-no interrogatives (e.g., Jed crammed the newspapers under the sofa in the trashcan) (cf. Peckenpaugh (2016)).

(56) *The current hypothesis:* The intuition may be extensible to PP attachment ambiguities that are syntactically disambiguated in addition to those that are pragmatically disambiguated (e.g., Had he planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?).

The goal of this study was to establish whether there is evidence for (56), and to explore the implications of those findings for possible explanations of (55).

4.1 Behavioral correlate for the 2016 intuition and the current hypothesis?

[14]: section title updated

Ultimately, no evidence has been found to support the current hypothesis, that the intuition can be extended to syntactically disambiguated sentences. Mixed-effect regression analyses were not able to detect statistical significance for the interaction between Speech Act and PP2 Status. This does not, of course, negate the intuition; it simply means that we have not yet found a behavior that can be said with certainty to correspond to that intuition. Future research should pursue the possibility that IRT (Inter-reading Time) is not the ideal measure for detecting any behavioral correlate of the processing difference for PP-attachment garden paths between interrogatives and declaratives.

4.2 On possible explanations for the intuition

This section looks at the evidence for and against some of the possible explanations for the intuition reported in Peckenpaugh (2016).

4.2.1 A prosodic account

The prosodic phrasings produced by participants in this study varied systematically by PP2 Status (Arg vs. Mod, where Arg is the garden path case). This is an interesting finding in and of itself, adding to the growing literature that shows a link between syntactic structure and prosodic phrasing. A possible explanation for the intuitive effect of interrogativity on parsing garden paths is provided in the work by Bader (1998). Bader demonstrates that it is easier to recover from[JDF] a failed[JDF] parse that "behave[s] alike prosodically" to a given failed parse, because the reanalysis does not require prosodic reconstruction and only the syntax needs to be repaired. In the case of the 2016 intuition~~this study is concerned with~~[JDF], this would mean that if sentences were more prosodically[JDF] similar across the Arg vs. Mod PP2 Status in the interrogative than in the declarative, the intuited~~ive~~[JDF] reduction in difficulty of reanalysis would naturally follow. This assumes, of course, that the 2016 iuntuition is eventually confirmed, despite the inconclusive results of Peckenpaugh (2016).

The findings of the current~~this~~[JDF] study show that there is no one-to-one mapping between prosodic structure and the four sentence types test in this~~defined by the design of the~~[JDF] study (Q Arg, D Arg, Q Mod, D Mod). Rather, for each Sentence Type,[JDF] ~~there are~~[JDF] gradient differences in occurrence of~~for~~[JDF] each pattern for a given sentence type were observed[JDF]. Recall that a break after the direct object is referred to as the object break (OBJ), and one after PP1 is referred to as the PP1 break (PP1).[JDF] Table **??** shows the distribution of the possible ~~simple~~[JDF] break patterns (PP2 break only, OBJ break only, or both breaks, with the negligible number of cases of neither break omitted) as a function of sentence type for Reading 2 only. These data are a subset of the data reported in Section 3.2.2 Table 3.5.

While there is a larger drop in the number of utterances with both breaks from Arg