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

Tyler J. Peckenpaugh

2019-08-26

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 of words (which lacks inherent structure). Following the models of parsing developed by, e.g., Kimball (1973), Frazier & Fodor (1978), and Frazier & Clifton (1996), this study assumes that parsing is done online , (i.e., during listening to or reading the word string) with the aim that most material is 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 temporarily ambiguous phrase in the input, which becomes apparent when subsequent material is encountered which cannot be incorporated into the resulting structure. This sort of parsing crash is called a ”garden path.” When it happens, the parser must reanalyze the material that had so far been processed, in order to arrive at a structure that can accommodate both the new and the old material grammatically.

In short: ”Garden path” effects occur when a temporarily ambiguous sentence resolves in such a way that the structure initially preferred by the parser is incompatible with how the sentence actually continues. These parsing errors have traditionally been attributed to structurally-driven parsing preferences (Frazier, 1979; Frazier & Fodor, 1978; Kimball, 1973) which ignore semantic

content on the first pass. Frazier (1979) formulates several of these structural preferences, 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 (Frazier, 1979, p. 20)

Because these strategies ignore semantic and pragmatic plausibility and the parser typically does not know what material might occur further on in the word string, mis-parses at temporarily ambiguous regions can occur. Minimal Attachment (MA) is important to the present study and will be revisited later on. 1.2.

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) in which *the horse* is modified by a reduced relative clause *raced past the barn*. Then *fell* can be incorporated as the matrix verb of the sentence, with a structure comparable to, *The horse (that was) raced past the*

barn was hungry.

The study reported in this dissertation is concerned with sentences that contain a temporarily ambiguous prepositional phrase (PP1), followed by another (PP2) which causes the initial parse to crash. Specifically, it is expected that PP1 in an example such as (4) will initially be interpreted as the goal of *cram*, but that parse will fail when it is realized that PP2 (*in his briefcase*) cannot plausibly modify *drawer*. Instead, PP1 will have to modify *paperwork* so that PP2 can be the goal argument of *cram*.

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

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

This contrasts with the sentence in (5), where PP2 can plausibly modify *drawer* and so the parse where the PP1 *in the drawer* is the goal argument is acceptable, and *of his filing cabinet* is incorporated as a modifier within it.

1.1 Motivations for the current study

The current study was initially motivated by a phenomenon discovered by Janet Dean Fodor and Dianne Bradley, and originally reported in Peckenpaugh (2016). That observation was 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 as a declarative.

(6) He had planned to cram the paperwork in the drawer in his briefcase.

(7) Had he planned to cram the paperwork in the drawer in his briefcase?

Peckenpaugh (2016) attempted to find a behavioral correlate of this intuition by examining variation in reading time for sentences like (6) and (7), but the results were inconclusive. The current study continues that line of research by testing similar sentences, while also attempting to control certain factors that may have led to Peckenpaugh (2016)'s inconclusive results. One such

concern is that (6) relies on the practical implausibility of a drawer within a briefcase, i.e., on real world knowledge, in order to disambiguate the appropriate attachment sites for PP1 and PP2. Real world knowledge and beliefs about what is or is not plausible likely vary between speakers and so may not always be effective as a trigger for reanalysis. The current study makes use of carefully constructed sentences (the criteria by which they were constructed are detailed in Section 2.1) which do not rely on plausibility or pragmatics to disambiguate the PP attachment sites, but instead make use of what will be referred to as syntactic disambiguation by including a PP2 that cannot grammatically be incorporated into PP1. This is illustrated in (8) and (9). Syntactic disambiguation is the term chosen to refer to the sort of disambiguation used for the current study, although various terms could be used, e.g., lexical or semantic disambiguation; it hinges on the lexical identity of the preposition that heads PP2, i.e., *into* instead of *in* and what syntactic position or semantic role such a PP can be assigned.

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

(9) Had he planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?

The change from pragmatic disambiguation in (6) and (7) to syntactic disambiguation in (8) and (9) creates an important distinction between the hypothesis tested in Peckenpaugh (2016) and the hypothesis to be tested in the current study . It cannot be assumed that the intuition about (6) and (7) necessarily extends to cases like (8)and (9), and so one of the questions the current study addresses is whether the greater ease of processing a question than a declarative 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 a clear focus on this small but potentially important difference, a convention will be 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 possibility that the intuition extends to syntactically disambiguated cases is referred to as the Current Hypothesis.

(10) *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., *The nanny sat the cranky little boy on the stroller on the swing*, vs., *Did the nanny seat the cranky little boy on the stroller on the swing?*).

(11) *The Current Hypothesis:* The 2016 Intuition may be extensible to PP attachment ambiguities that are syntactically disambiguated in addition to those that are pragmatically disambiguated (e.g.,*He had planned to cram the paperwork in the drawer into his briefcase*, vs., *Had he planned to cram the paperwork in the drawer into his briefcase?*).

In addition to exploring the possible extensibility of the 2016 intuition, the current study addresses 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 declaratives. Because there are minimal differences between the polar question and declarative versions of these sentences, it seems likely that the cause lies in one of two domains: either 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? the reported study seeks to answer this, but while the recordings collected provide some insight, more work is likely needed to provide aa fully satisfactory answer.

Likewise, the latter possibility raises the question: exactly what are the semantic and pragmatic differences between a polar question and its declarative counterpart? This question has been approached in the theoretical literature (see, e.g. Fiengo (2007)). Nevertheless, it remains to be determined how those properties could lead to an easier or more difficult parsing process.}

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 prove viable.

The example in (12) shows a pragmatically disambiguated sentence with an argument-PP2.

(12) Jed crammed the newspapers under the sofa in the wastebasket.

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]]

"#" indicates a structure with an implausible reading

The initial parse is expected to be (12 a) because of a bias (due to Minimal Attachment, or some variation thereof, see (1) above), 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] with PP1 denoting the goal, which leaves nowhere for the second PP to attach but as a modifier of the noun phrase (NP) inside PP1 ([$_{PP1}$ *under* [$_{NP}$ *the sofa* [$_{PP2}$ *in the wastebasket*]]]). This initial parse (12 a) is pragmatically implausible, since sofas are generally not found inside wastebaskets. Structural reanalysis is required to bring about the only plausible parse (12 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 [$_{VP}$ V [$_{NP}$ N PP1] PP2], i.e., where the newspapers under the sofa are being crammed into the wastebasket.

For clarity in what follows, the current study categorizes the sentence types being discussed into two groups based on the status of the PP2 they contain: (a) cases where PP2 is an argument of the verb are referred to as Arg-type sentences, and (b) cases where PP2 is a modifier of the preceding noun are called Mod-type sentences ; see (13). As noted, it is the Arg sentences that are generally reported to be garden paths, because PP2 must fill the goal role that PP1 is expected to have filled as just discussed. Mod sentences are not likely to result in a garden path, because PP2 can modify the

NP within PP1 to become part of the goal, and therefore need not disrupt PP1's attachment as the goal.

(13) **PP2 attachment status**

(a) *PP2 Argument* (Arg)

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

(b) *PP2 Modifier* (Mod)

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

In order for the differing parsing process for (13 a) and (13 b) to be explained by a strictly structure based model of parsing, certain assumptions have to be made about the syntax. A simple way to achieve it 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 such as *cram*. In this case, in order to avoid postulating extra nodes that would be required for PP1 to be a modifier of the object NP, Minimal Attachment dictates that PP1 should be assumed to fill the verb's goal argument slot. However, this is not how current syntactic theory assumes the structure to be, as three-way branching is proscribed (see e.g., Chomsky (2014)).

The most pronounced structural difference between the two structures is that in the sentence with a modifier-PP2 (Mod) the major disjuncture comes fairly early, just after the object NP *the newspapers*. For the sentences with argument-PP2s (Arg), the major disjuncture is later, just after PP1 (*under the sofa*).

It bears mentioning 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 a verb with multiple internal arguments incorporating each of those arguments as a sister (i.e., a ternary branching structure as in 1.1).

Within 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

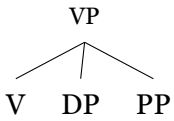

Figure 1.1: Illustrative syntactic tree of a ternary-branching VP.

Figures 1.2 and 1.3 below[1], so it becomes less clear that the VP attachment site for PP1 actually creates fewer nodes than the lower NP attachment site, as per Minimal Attachment. Nonetheless, studies by, e.g., Rayner, Carlson, & Frazier (1983), Clifton, Speer, & Abney (1991), and others, show that a preference for VP attachment of a PP in contexts such as the one the current study addresses exists, whether that preference is best explained by Minimal Attachment, or a preference for arguments over non-arguments, or some other mechanism.

Setting the particularities of syntactic theory aside now„ an appeal can be made to a psycho-linguistic distinction added to parsing theory in *Construal* (Frazier & Clifton, 1996): that of primary vs. non-primary relations.

(14) ”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)

This additional contrast is independently motivated: Frazier and Clifton illustrate this by way of relative clause (RC) attachment in constructions like (15).

(15) The journalist interviewed [$_{NP1}$ The daughter$_i$] of [$_{NP2}$ the colonel$_j$] [$_{RC}$ who$_{i/j}$ had an accident] (Frazier & Clifton, 1996, p. 71).

The RC in (15) can modify either NP1, *the daughter*, or NP2 *the colonel.* Late Closure(see (2)) predicts a consistent preference for local attachment of the RC in (15), i.e., the structure where the RC modifies NP2. This is because Late Closure dictates that the NP2 phrase not be closed until

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

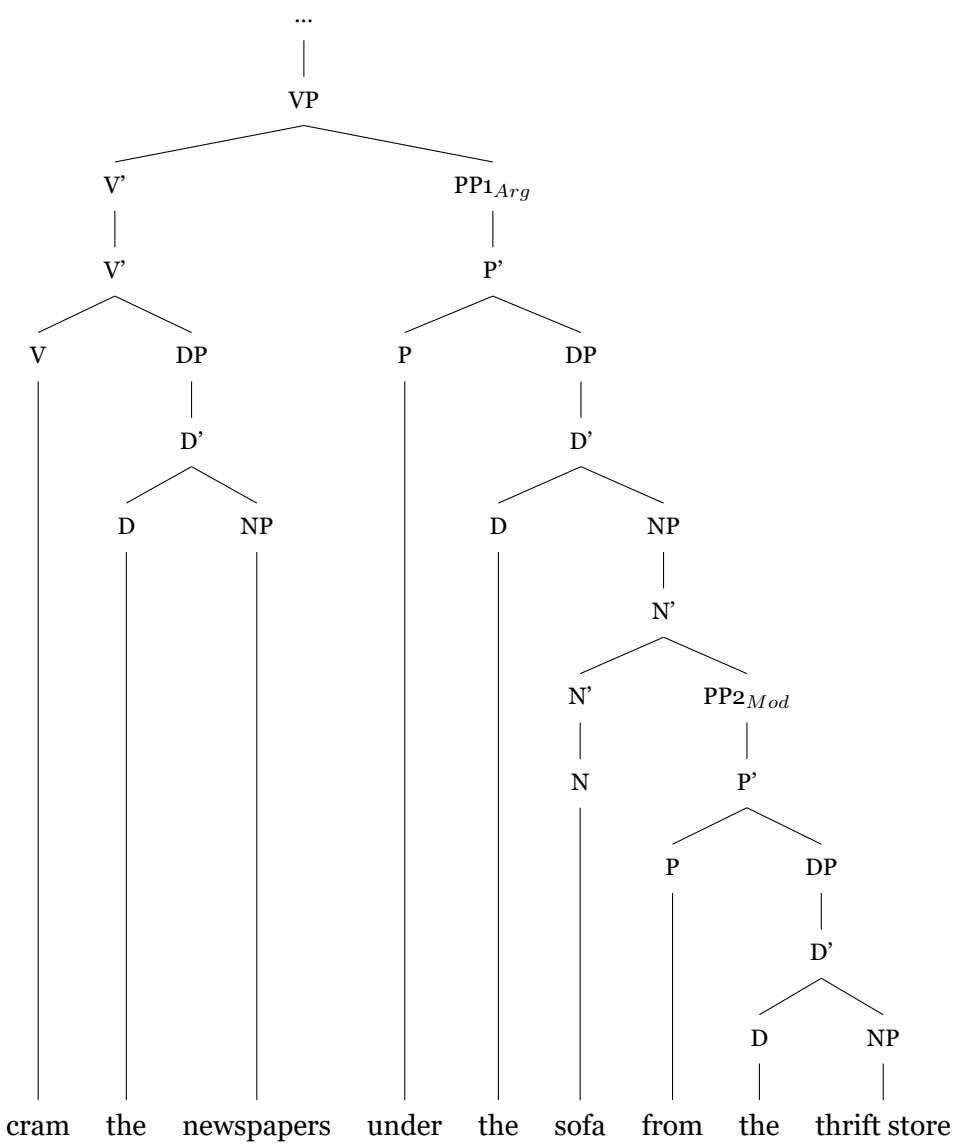

Figure 1.2: Syntactic tree of an illustrative example sentence with an ambiguous PP1 and a modifier-PP2 (Mod).

material that cannot, for one reason or another, be incorporated into that phrase is encountered. Since the RC can modify NP2, Late Closure predicts that it will. Instead, what a number of empirical studies (e.g., Clifton, 1988; Cuetos & Mitchell, 1988) find is a pattern where the preferred structure depends on the relationship between NP1 and NP2. Frazier & Clifton (1996) describe five categories of relationship, and a gradient of preferred RC attachment, from NP1 preference to NP2 preference.

(16) **RC Attachment by NP1-NP2 relation** (Frazier & Clifton, 1996, p. 31)

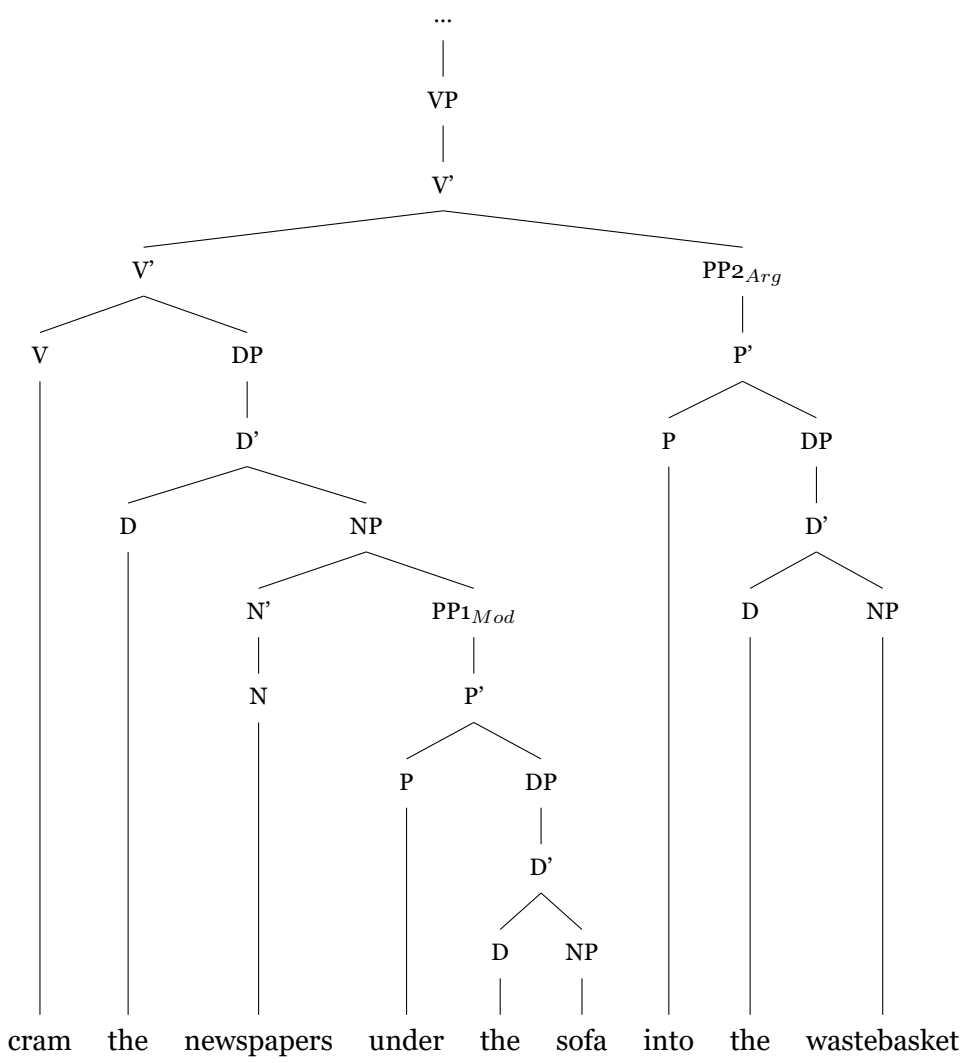

Figure 1.3: Syntactic tree of an illustrative example sentence with an ambiguous PP1 and an argument-PP2 (Arg).

a) *Material*

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

b) *Quantity*

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

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

   The son of the woman [$_{\text{RC}}$ 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 that (16 a-b) type configurations favor NP1 RC attachment, (16 e) type configurations favor NP2 RC attachment, while (16 c-d) are intermediate. They argue that this gradient cannot be readily explained by structural parsing, and instead they make use of a mechanism they call 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 (16), the last theta assigner is NP2, and its domain extends up to the DP that contains NP1. Association is a later parsing decision that allows the syntactic structure to be decided after semantic information becomes available: the RC can ultimately modify whichever member of the thematic domain is appropriate.

The factor that determines when association or structural parsing is appropriate is the distinction between primary vs. non-primary relations. Frazier and Clifton's formalization of this distinction was previously mentioned as (14) above.

RC attachment undergoes association because relative clauses are by definition modifiers, and the relationship between a modifier and whatever is modified is a non-primary relation Returning now to the PP-attachment construction that this study is concerned with, the argument vs. modifier distinction is precisely what distinguishes the two possible roles of PP2 shown in (13), repeated here as (17).

(17) **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]].

The infinitival clause headed by *cram* is a primary phrase, and so its obligatory constituents hold primary relationships with *cram*. As such, association is not available as a mechanism and structural parsing must occur, without semantic information being available. Thus, PP1 is initially interpreted as the goal of *cram*. Then, in the case of (17 a), when PP2 is encountered, PP1 must be removed from its syntactic position and be reanalyzed as a modifier of the object NP in order to allow PP2 to fill the goal argument position. In the case of (17 b), no reanalysis is necessary because PP2 can modify *the drawer* and so does not need to dislodge PP1.

1.3 Interrogativity

This section returns to the questions raised by the 2016 Intuition discussed in Section 1.1 about why certain garden paths of the sort just discussed might appear to be easier to parse in interrogative constructions than in otherwise similar declarative ones (see examples (18) and (19)).

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

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

What exactly differs between (18) and (19)? Syntactically, very little: the position of the subject *he* and the auxiliary *had* have been reversed, but the words that follow are identical.

The semantic, or perhaps more accurately pragmatic, differences between (18) and (19) lie with the presuppositions the sentences carry with them, which may vary with the placement of focus, marked phonologically in spoken language. The exact phonetic properties of focus, and the semantic and syntactic consequences of it, are widely studied and debated. For an excellent overview of the topic, see Ladd (2008), pp.213-23. The assumptions made about focus in this dissertation are about English only, and are more or less compatible with a Focus-to-Accent model as described by Ladd, which distinguishes ”the semantic/pragmatic notion ’focus’ from the phonetic/phonological notion ’accent’ and – crucially – [...] allows focus to apply to portions of utterances larger than individual words” (Ladd, 2008, p. 217). This results in sentential stress

being split into two related but independent components: where focus lies within the sentence, and how it is conveyed with regard to the location of accent.

The declarative in (18) 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 (19) are different from those of (18): for instance, a yes/no question additionally presupposes that the listener does or may know the answer to the question.Further presuppositions may exist, depending on where the focus lies within the sentence.

Focus in a declarative like (18) is typically wide, i.e., no element is having special attention called to it. A polar question like (19), however, will often receive narrower focus on one word or phrase, so that when uttered, that word or phrase 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 in the sentence (subject, auxiliary verb, matrix verb, infinitival verb, object NP, or the NP of either PP1 or PP2) or any part of one of these. Words in all capital letters indicates verbal emphasis signifying focus.

(20) HAD he planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?
(21) Had HE planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?
(22) Had he PLANNED to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?
(23) Had he planned to CRAM the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?
(24) Had he planned to cram the PAPERWORK [$_{PP1}$ in the drawer] [$_{PP2}$ into his briefcase]?
(25) Had he planned to cram the paperwork [$_{PP1}$ in the DRAWER] [$_{PP2}$ into his briefcase]?
(26) Had he planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into HIS briefcase]?
(27) Had he planned to cram the paperwork [$_{PP1}$ in the drawer] [$_{PP2}$ into his BRIEFCASE]?

In (20), with focus on the auxiliary, the question encompasses the entire proposition, and asks 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 (21), 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*? In (25), with focus on *drawer*, the question is instead about which

paperwork this is all happening to: the paperwork that is 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 (18) and (19) might be the source of the 2016 Intuition that (19) is easier to comprehend than (18), but it is not entirely clear why, and it is not the only possibility. Another significant difference between the two sentences is the rhythm and intonational melody (which, taken together, make up the prosody). While dialects of English may differ prosodically, there is typically a difference in melody between a declarative and a 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 be shown to correlate with a difference in processing difficulty in syntactically disambiguate declarative and interrogative PP-attachment garden paths, and by extension lend insight into the 2016 Intuition.

1.4 Prosody of questions vs. declaratives

In pursuing the possibility that it is the prosody of polar interrogatives which creates the 2016 Intuition that motivated this study it is important to consider the details of question intonation actually sounds like. It is generally agreed that in American English, the intonation of a polar question has exhibits a final rise. 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 their ToBI notation, a tone T is coded either L for low or H for high; T* is anchored to the stressed

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

syllable, and T- and T% are boundary tones (intermediate phrase boundary and intonational phrase boundary respectively). See *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 categorize 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 total data showed any sort of falling contour. According to the authors' analyses, final high-rise 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 in (28), regardless of whether the final PP is a Mod or an Arg.

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

The need to prepare for that final 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 possibility is revisited and more fully explained in Section 4.2.1.

The prosodic structures observed in the data collected for the current study are discussed in 3.2. For discussion of what constitutes a prosodic boundary, see e.g., Streeter (1978) and Salverda, Dahan, & McQueen (2003).

1.5 Evidence that prosody can affect syntactic parsing

A number of studies have shown that in listening to speech, prosodic cues 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 temporary ambiguity which could result in a garden path. They recorded speakers saying sentences with natural prosody, such as the following pair (not bracketed in presentation to the speakers):

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

(30) [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 which favored the garden path. The initial fragment of each (the portion from the beginning of the sentence to the word *house* in (29) and (30)) was then presented to participants who 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 visually presented word was compatible with the structure indicated by the prosodic cues. The results of this study, as well as a growing body of literature, suggest that prosodic information can indeed be used by the parser in making processing decisions.

Fodor (2002) invoked prosody in an explanation of differing relative clause attachment preferences across languages first pointed out by Frazier & Clifton (1996) and discussed above in Section 1.2. This concerns sentences such as English (31) and Spanish (32):

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

(32) Alguien disparó contra la criada de la actriz que estaba en el balcón.

The relative clause in (31) *who was on the balcony*, can attach either locally (modifying N2), making it *the actress* who was *on the balcony*, or non-locally (so that it is *the servant* who was *on the balcony*. While Late Closure predicts local/low attachment in these structures, Cuetos & Mitchell (1988) found a 60% preference for local/low attachment for the English materials, but only a 40% preference for local/low attachment in Spanish . In apparent violation of the general preference for local/low attachment, some languages, like Spanish (and French and Russian, but not Romanian or Brazilian Portuguese, so this is not a general feature exclusive to Romance languages), prefer to attach relative clauses higher, while others more often obey Late Closure as in English (e.g., Swedish and Egyptian Arabic). Interestingly, the non-local preference in languages like Spanish is weakened

in cases where the ambiguous RC is short (one prosodic word), which is compatible with its being attributable to prosodic phrasing; see below). Fodor (2002) maintains that these tendencies exist both in listening to spoken sentences, with and without explicit prosody and in silent reading.

Other researchers, e.g., Maynell (1999), have shown that the presence or absence of a prosodic break before an RC encourages high or low attachment respectively. Fodor leverages this in order to explain the differences in RC attachment site tendency between languages. She argues that the cross-language differences can be neatly explained 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 constraint-based approach. Consider Selkirk's (1986) alignment constraints:

(33) 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 this general idea for English.

(34) **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).

The 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 aligns with the left edge of the RC in examples like (32), *Align-XP/L* is postulated to be ranked above *Align-XP/R* in languages like Spanish that prefer high attachment since a prosodic break in that place has been shown to encourage a high attachment interpretation (Maynell, 1999). In languages where low RC attachment is preferred, *Align-XP/R* can be postulated to be ranked higher, and thus a prosodic break is more likely to occur after the RC than before it.

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

(35) **BinMin(ϕ)**

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

If , in Optimality Theoretic (Prince & Smolensky, 1993) terms, 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, because it needs more material in order to achieve a length of at least two phonological words, and so must appropriate some from the preceding NP. 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, as they have enough phonological content to stand alone.

1.6 Predictions for the current study

The experimental study presented here addresses a number of issues. First: is syntactic phrase attachment in any way encoded in the speech signal? I hypothesize, following e.g., Schafer, Speer, Warren, & White (2000) as noted above that it is, and therefore that prosody can be used to diagnose attachment site. Consider the basic configuration in Figures 1.4 and 1.5 below.

The predictions made here rely on there being an ideal prosodic structure for the sentences tested

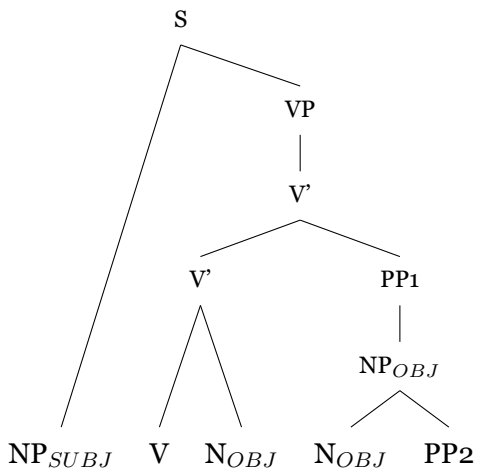

Figure 1.4: Illustrative syntactic tree of the basic configuration for Mod cases.

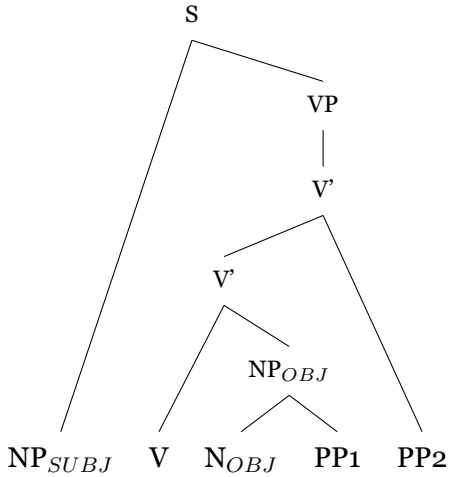

Figure 1.5: Illustrative syntactic tree of the basic configuration for Arg cases.

to compare them with. Following Fodor (2002), a major disruption to the syntax like the one after PP1, as in Figure 1.5, should be marked by a prosodic boundary between PP1 and PP2. On the other hand, modifier attachment of PP2 will lack any substantial boundary marking in that position unless it is necessitated for length reasons; instead, a break after the object is expected. These expectations are shown in (36) and (37), where "|" a less prominent or no break and "||" represents a linguistically motivated prosodic break.

(36) ... stick the letter || in the mailbox | of the proper stack
OBJ Break PP1 Break
Direct object PP1 vice president.

(37) ... stick the letter | in the mailbox || onto the proper stack.
OBJ Break PP1 Break
Direct object PP1 PP2

An issue that is addressed is 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 semantically implausible parses should be more frequent in Reading 1 of a sentence with argument-PP2 than in a second reading of the same sentence.

(38) *Hypothesis 1* A first reading of a sentence where PP2 is a goal argument (Arg) will exhibit less

natural prosody (more hesitation at and within the PP2 region) than:

a. A first reading of a sentence where PP2 is a modifier (Mod)
b. A second reading of a sentence where PP2 is a goal argument (Arg)

(39) *Hypothesis 2* A first reading of a sentence where PP2 is a goal argument (Arg) 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 of that sentence will more often be pronounced with the prosodic structure that represents the intended parse (argument attachment of PP2).

Hypotheses 1 and *2* together make a third prediction: readers should struggle more on the cold reading of a sentence with an argument PP2 (Arg) to obtain a plausible structure, and thus the appropriate prosody, than on a second reading.

(40) *Hypothesis 3* 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.

Finally, the question of whether the 2016 Intuition described in Section 1.1 can be shown to extend to the syntactically disambiguated sentences tested in the present study is formalized as *hypothesis 4*. This hypothesis presumes that IRT reflects processing difficulty.

(41) *Hypothesis 4* The inter-reading time (IRT) will be longer for Arg sentences that are declarative than for:

a. Arg sentences that are interrogative
b. Mod sentence that are interrogative or declarative

Note that these hypotheses are not applicable in cases where the reader fails to successfully and fluently produce the sentence.

These hypotheses are returned to in Chapter 3. The findings from the current study that will be

presented in what follows do not successfully settle all of the issues raised here, but it is hoped that they will help guide future work in directions that may do so.

Chapter 2

Methodology

This chapter describes the methodology employed for the study reported in Chapter 3. 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 a visually presented sentence 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) under the assumption that in order to pronounce a sentence with the optimal prosody, it's necessary for the speaker to understand the sentence. 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 the prosodic grouping that readers used, suggesting that comprehension was improved on the second reading.

While the questions being addressed in the current study are different from those of Fodor et al. (2019), it is still concerned with the prosody that is produced, as well as the degree of difficulty the reader experiences in parsing a sentence in order to read it aloud. This experimental paradigm

eliminates the possible uncertainty of not knowing whether a given pronunciation represents a naive or considered attempt to read a sentence aloud.

2.1 Materials

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

2.1.1 Experimental items

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 must be a modifier (Mod) of the preceding phrase (PP1). A full list of experimental items is available in Appendix A.

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

Version	Sentence
D Arg	He had decided to stick the large check in the envelope into her wallet.
Q Arg	Had he decided to stick the large check in the envelope into her wallet?
D Mod	He had decided to stick the large check in the envelope from her church.
Q Mod	Had he decided to stick the large check in the envelope from her church?

As previously mentioned, the current study was motivated by an observation first reported in Peckenpaugh (2016). The experimental stimuli used in the current study were based on the materials from that study 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 (42). Note that the material starting with the infinitival verb, e.g., *cram* until the end of sentence will be referred to as the construction throughout this and later chapters, as labeled in the example below.

(42)

Introductory material			Construction			
Subject	Auxiliary	Matrix verb	Infinitival verb	Object	PP1	PP2
Order shown for D versions (reversed in Q)					always ambig-uous	Disam-big-uation of PP1

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 rationale for the use of these mental state verbs are discussed later in this section.

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 postulate PP1 to be the goal argument due to the primary status of arguments, predicts that (all and only) Arg versions of a quadruple require will 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* in 8 cases, *on* in 7 cases, and in one case, *under*.

A benefit of using a complex verb cluster (auxiliary + matrix participle + infinitive; see **??** above) rather than a single verb[1] was that the differences between declarative and interrogative versions of a quadruple were isolated to the left extremity of the introductory material of the sentence: i.e., only the position of the subject and the auxiliary were affected, meaning that the construction itself and several words prior to it were completely untouched by the Speech Act manipulation. The construction is in a sense buffered from changes triggered by Speech Act manipulation, and is only effected by the PP2 Status manipulation.

The purpose of including introductory matrix verbs (e.g., *intended*) 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 *intended*: rather than asking about facts that we already seem to know, we are instead asking about an actor's mental state with 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

[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.

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: *cram*, *put*, *set*, or *stick*. Verbs that optionally take a goal (e.g., *hide*) 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 *hide* and *cram* shown in (43) and (44):

(43) **Optional goal** (*hide*)

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

✓ The gangsters had hidden the shotguns.

(44) **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 (43), can take a goal, but can also be used without one. A verb like *put*, on the other hand, as in (44), really must have a goal even if not fully detailed (e.g. *put down*). 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*. While certain constructions containing these verbs do exist where no goal is needed (e.g., *he student had needed to cram all night*; *the narrator had set the scene*; *the disgruntled worker decided to stick with the program*; etc.), it is arguable whether such instances should be considered the same lexical item as the ones being used here, and in any case all experimental items used in the current study would be rendered ungrammatical by the omission of the goal argument, as with the example in (44).

Another important consideration was ensuring that the Arg versions had a PP2 which definitively disambiguated the attachment site of PP1 from the goal argument role to being a modifier of the

object; i.e., that reanalysis was forced.

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

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

In (45), PP2 is implausible as a modifier of *rocking horse*, but not strictly impossible syntactically, and the sentence is grammatical with PP2 modifying it. On the other hand, the use of *onto* in (46) completely disallows the modifier interpretation of PP2 at the syntactic level: a PP headed by *onto* cannot grammatically modify the preceding NP.

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 inconsistency in the results 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 long, 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, PP2 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, which is 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 in the +PP filler items was the same (e.g., there were 5 cases where PP2 was an argument), except there were 6 that modified the NP embedded in PP1 instead of 6 that modified the object. The purpose of this decision was to avoid any discrepancy that might result from any difference in processing difficulty of certain configurations of attachment sites for the PPs. An even distribution of attachment site for both PPs ensures that any analysis over the filler items should not be unduly impacted by PP attachment site when that is not what the analysis is concerned with, and it makes available an analysis of those attachment sites, although that analysis is not pursued here. A full list of fillers is available in Appendix B.

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 verbs used for the experimental items (*decide*, *intend*, *plan*, and *want*), or one of 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

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?

a participant would see 6 instances each of *decide*, *intend*, *plan*, and *want*, (i.e., twice in experimental items and 4 times in filler items) 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 in order 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

Sentence 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 than 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 current study 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 via a software system designed for university participant pools. Students saw a recruitment notice on the system website (see Appendix C), 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 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,

[3] IRB approval number: 2018-0072

[4] Instructions were also provided verbally and via printout, see Appendix D for the latter.

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 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 D) before signing a consent form. Participants then sat at the computer and were again walked through instructions before the first practice item was presented. The instructions made clear that the task consisted only of reading each sentence twice in succession, under different timing conditions as specified.

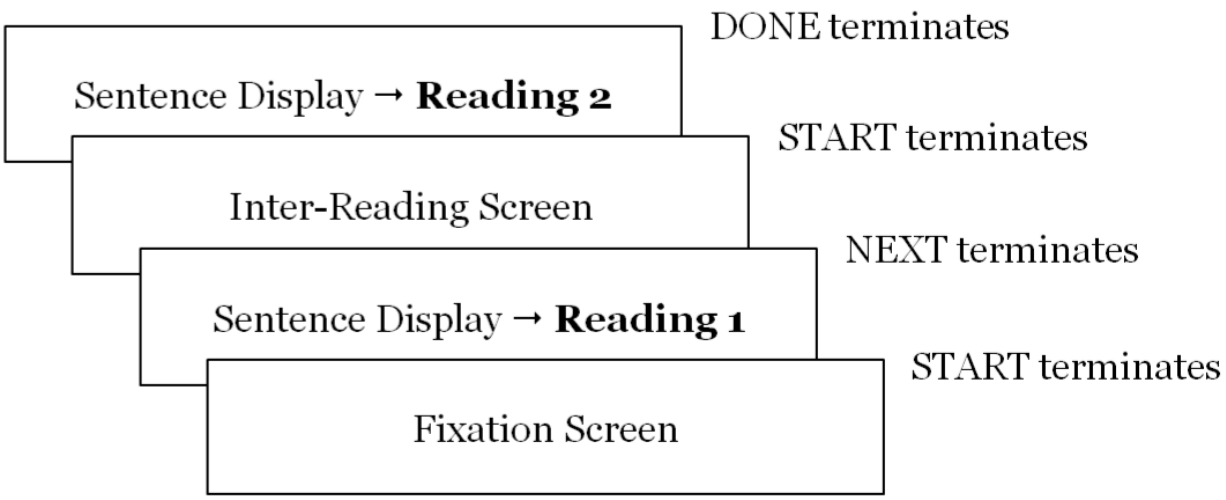

Figure 2.1: Diagram of 4-screen sequence presented for each item, showing the key presses triggering movement between successive screens.

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.

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. 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 of the Double Reading Procedure is that it allows for certain assumptions to be made about Reading 2 for the purposes of analysis that otherwise would be unclear to the researcher: Reading 2 certainly represents a considered reading of the sentence. The reader not only has necessarily read the sentence and (heard it read) in producing Reading 1, but has had ample time to examine the sentence . This means Reading 2 can plausibly be thought to represent a considered prosodic and syntactic 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 228 wpm, 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 R1 delay was 1s which would allow a fast reader time to read very few words. In fact, the window is even shorter, because in addition to just reading, the subject is also performing several other cognitive processes (e.g., visual processing, lexical access, issuing motor commands, etc.). For most recordings, therefore, the utterance of Reading 1 should contain within it any behavioral reflex of whatever online parsing difficulty the reader has.

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 the time course of reading is thoroughly summarized in Rayner, Pollatsek, Ashby, & Clifton (2012). They describe reading as consisting of a series of fixations, during each of which 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, it can be inferred 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 here, 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. The mutilation of text outside a window of clear text creates a so-called moving window around the reader's point of fixation is created. By manipulating the size of this window, it has been found that reading speed is maximized when about 15 characters to either side of the fixation site is accessible to the reader (it turns out this is actually asymmetric, and the window need only extend far as the start of the currently fixated word in the direction of what has already been read, i.e., to the left for English readers).

In a study by McConkie & Rayner (1975), 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 other 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 occurred when the originalharacters outside the moving window 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, McConkie and Rayner 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 from the current study is displayed in (47), 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.

(47)

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

Table 2.3 presents these distances averaged across items all experimental items. Note that these values do not vary across condition, because the initial fixation is locatedcounting starts after both the subject and auxiliary verb, and the counts end before PP2. The only changes across versions are subject-auxiliary inversion and the content of PP2 which fall outside the string of material over which these counts were calculated.

Table 2.3: Distance in characters from fixation to disambiguation of experimental items for the current study.

	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. Recall that word boundary information is available only 15 to 18 characters to the right of fixation, thus the 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 fourth fixation point (the head noun within PP1) to the disambiguating region, PP2, 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 (47) 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 factor 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, that 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 position for a break to

hesitate due to the garden path effect of discovering the disambiguating PP2. This property of oral reading calls into question the status of OBJ breaks reported in 3 with regard to whether they are linguistically motivated, or represent the processing difficulty experienced when PP2 is finally observed by the reader.

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, because the experiment was self-paced, and the participant might spend time studying the sentence after their first reading but before advancing to the next frame. For this reason, 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 is 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 remaining recordings 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 phonation, with breaks in phonation of less than 1s not breaking contiguity, because 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

[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)

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 possibly very slow readings of the admittedly difficult items being tested in the current study.

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 recordings. 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).

2.8 Prosodic judgments

A trained linguist informant naive to the research being conducted listened to the 978 recordings of experimental items (note that 46 recordings were missing or omitted, as discussed in Section 3.1) and reported the presence or absence of breaks in certain regions of each sentence, as well as

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

several other judgments (e.g., the relative strength of these breaks, whether or not the reader struggled, and whether or not the reader used final-rise). Analyses of some of these judgments (e.g., the presence or absence of the OBJ and PP1 break) are reported in Chapter 3, while others lacked interesting results and were not reported (e.g., the presence or absence of the V break). The rater 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 (48), as well as full plain-text lists of all items.

(48)

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

The rater was asked to report on whether or not she heard a prosodic boundary directly following the region labeled **OBJ**, and directly following the region labeled **PP1**. Instruction regarding how to define prosodic break was provided by way of the following prose:

> 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) in the same utterance.

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 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 E.

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, also naive to the purpose of the research, 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.

Table 2.5: Percent agreement between the original ratings and the second rater (inter-rater) or the second rating by the original rater (intra-rater).

	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

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.

Chapter 3

Results and discussion

This chapter 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 by a participant considering a sentence between readings, which will be called 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 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 2 readings x 48 items, 16 experimental and 32 filler) 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 is a property of pairs of utterances (Reading 1 and Reading 2) rather than single recordings, the database for response timing took in 489 data points.

[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.

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

	D	Q
Mod	246	248
Arg	244	240

3.2 Prosodic break patterns

This 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. These data are 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 infinitival 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 (49).

(49)

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

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. A second linguist provided judgements over a subset of the recordings, and a comparison between their judgments is available in Section 2.8.1 above.

3.2.1 Individual break patterns

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 3.3: 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
Mod	77.2% (95)	76.6% (95)	84.6% (104)	72.6% (90)
Arg	57.4% (70)	56.7% (68)	73.0% (89)	74.2% (89)

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.

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
Mod	68.3% (84)	68.5% (85)	84 (68.3%)	83 (66.9%)
Arg	99.2% (121)	99.2% (119)	121 (99.2%)	117 (97.5%)

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 all subsequent reports of prosodic patterns.

There was very little difference across readings, with the following generalizations of the data shown in Table 3.5 holding for both Reading 1 and Reading 2. For Mod sentences the OBJ-only

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	31.7%	30.9%	0.8%	0.8%	31.1%	31.4%	0.8%	2.5%
Both	45.5%	46.3%	56.6%	55.8%	54.1%	43.0%	72.1%	71.7%
PP1 only	22.8%	22.8%	42.6%	43.3%	14.8%	25.6%	27.0%	25.8%

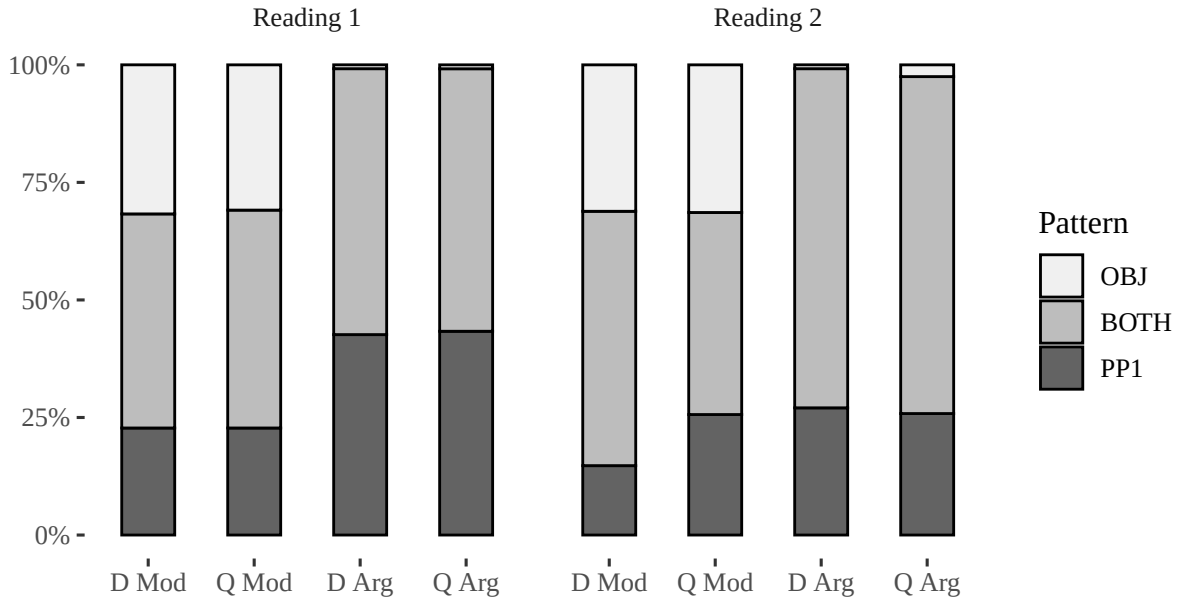

Figure 3.1: Break pattern as a function of sentence type and Reading.

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

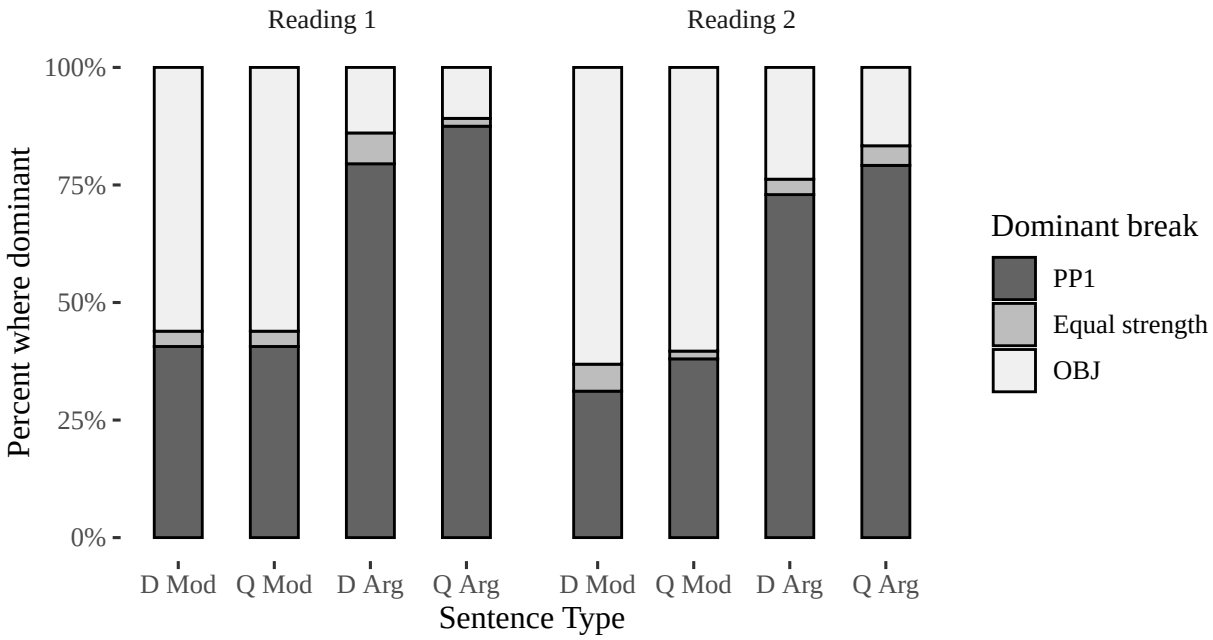

Figure 3.2: Percent break dominance occurence as a function of sentence type and Reading.

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).

For both the analysis where relative strength is ignored, and the analysis of break dominance, the data can be thought of as existing in of three bins: the PP1 (only or dominant) bin, the OBJ (only or dominant) bin, and a neutral (both breaks present or neither break dominant) bin between them. In Section 3.2.2, the neutral bin containing instances of both breaks occurring is robust. The break dominance analysis in Figure 3.2 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 cases of ”neither dominant” for the strength-sensitive data than of ”both breaks present” for theinstead of simple occurrence data.

Figure 3.2 clearly shows a robust effect of PP2 Status (Mod or Arg) on break dominance, and little to no impact of Reading or Speech Act (Q or D). A dominant break after PP1 is frequent for all Arg sentences, both declaratives and questions, and in both Reading 1 and Reading 2. It is less fredquent for Mod sentences in both Reading 1 and Reading 2. The higher relative frequency of PP1 break dominance in Arg sentences compared to Mod sentences is expected, since it represents a linguistically motivated prosodic break in the Arg cases. For the Mod sentences, there is no linguistic motivation for the PP1 break, so it makes sense for it to be less frequent and less dominant in Mod sentences. It is noteworthy and puzzling that the patterns do not differ greatly across the two readings, as one would expect difficulty getting the prosody right on the first try for the difficult sentences being tested in the current study.

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 model 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

[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.

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 3.10: 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 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.

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

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.

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 sentences 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.

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

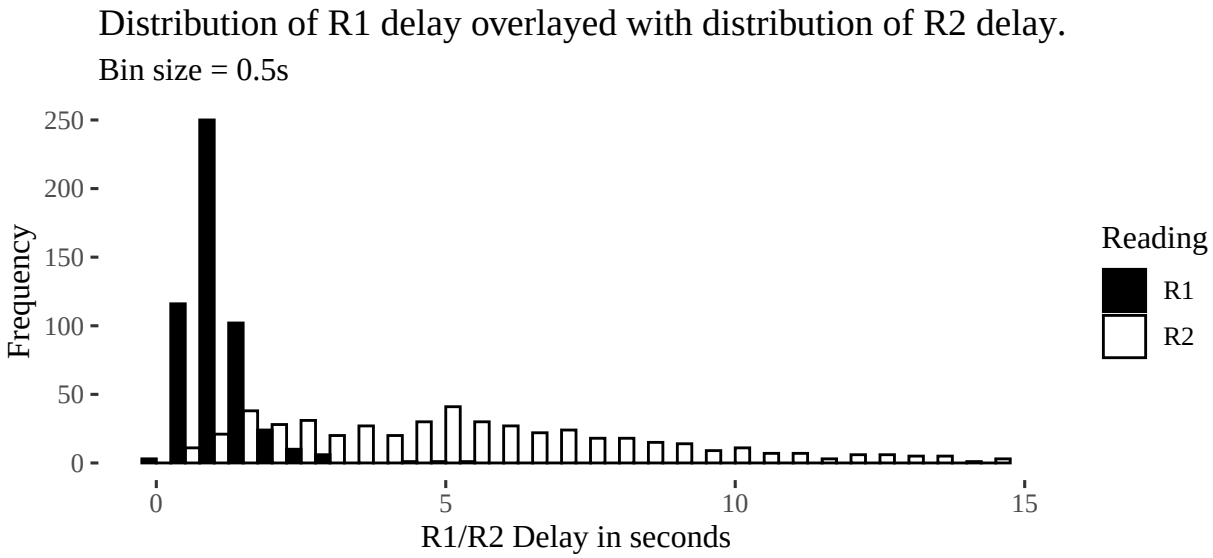

Figure 3.3: Distributions of R1 delay and R2 delay

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 ($\chi^2(1) = 4.05$, $p < 0.05$), 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 for Arg compared to Mod sentences for Reading 2 within the slow R1 delay category ($\chi^2(1) = 3.97$, $p < 0.05$), but comparisons across other factors represented in Table **??** 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 (PP1 or OBJ) than if R1 delay was fast. A possible explanation is that for recordings in the slow category readers were more prone to hesitation in general, or perhaps, contrary to instructions, were using both the delay time as well as the extra time created via hesitation to look ahead. The significant effect of

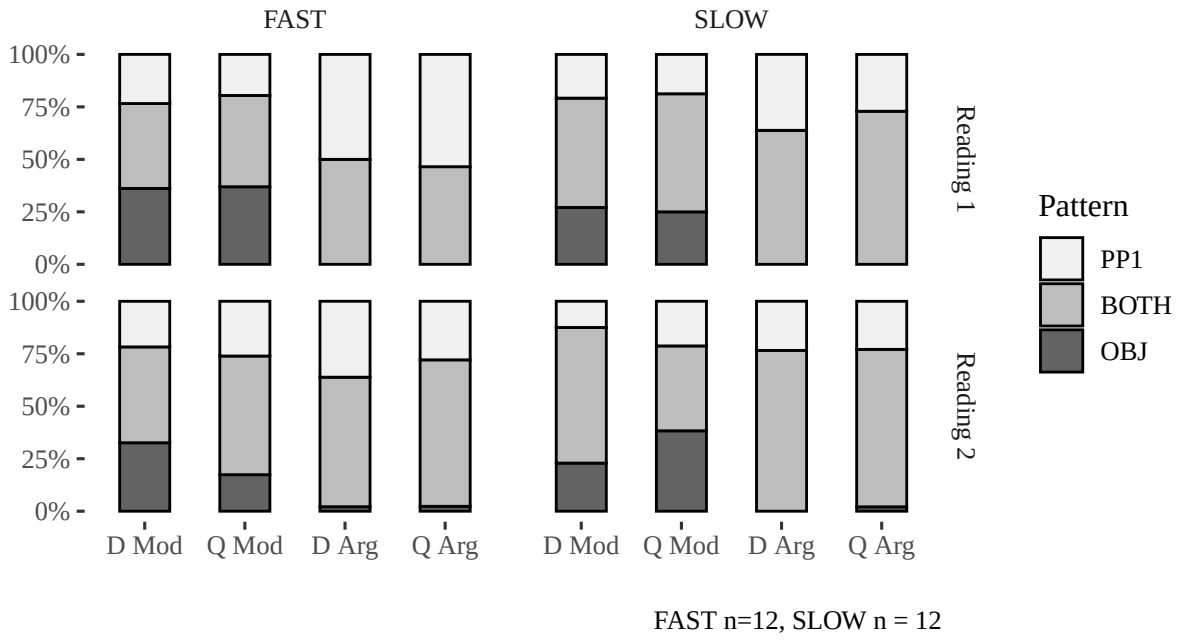

Figure 3.4: Plot of pattern proportions as a function of sentence type.

PP2 Status for the slow category in Reading 2 could be due to the reader not fully understanding the Arg sentences prior to Reading 2, thus increasing their likelihood of hesitation; i.e., where the fast category represents confident readers, the slow category represents less confident readers. As a result of the reader's increased difficulty in Reading 1 for the recordings in the slow category, the effect of processing difficulty is not limited to Reading 1 but in fact spills over into Reading 2. The slow category represents readers who are never able to fully comprehend the sentence, and thus hesitate more frequently.

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 (see Section 3.2.4.1). 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; conversely, 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.

Returning to the predictions discussed in 1.6, recall the contrast made there between linguistically motivated prosodic breaks and other breaks or pauses. The Arg and Mod sentences differ with regard to where the linguistically motivated breaks fall. These expectations are shown in (50) and (51), where "|" a less prominent or no break and "||" represents a linguistically motivated prosodic break.

(50) … stick the letter ||| in the mailbox | of the proper stack
OBJ Break PP1 Break
Direct object PP1 vice president.

(51) … stick the letter | in the mailbox || onto the proper stack.
OBJ Break PP1 Break
Direct object PP1 PP2

The break patterns observed in the data just discussed mostly do reflect the expectations laid out in (50) and (51): PP1 is the dominant break in a majority of Arg cases for both Readings, and OBJ is dominant in less than half of Arg cases for both Readings (see Figure 3.2 above). Conversely, for Mod cases, OBJ is the dominant break in a majority of cases and PP1 in less than half, across both Readings.

That Reading 2 is a significant predictor of both OBJ break dominance and PP1 break dominance (see Section **??**) supports, at least provisionally, hypotheses 1 and 2 from Section 1.6, repeated below.

(52) *Hypothesis 1*

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

a. A first reading of a sentence where PP2 is a modifier (Mod)
b. A second reading of a sentence where PP2 is a goal argument (Arg)

(53) *Hypothesis 2*

A first reading of a sentence where PP2 is a goal argument (Arg) 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 of that sentence will more often be pronounced with the prosodic structure that represents the intended parse (argument attachment of PP2).

The present study did not specifically attempt to assess whether a given reading represents more or less natural prosody for these constructions.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 one without as much preview. *Hypotheses 1-2* are supported only on the assumption that this is so.

(54) *Hypothesis 3* 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.

Hypothesis 3 is not supported by the evidence just reported: there is no statistically significant interaction between Speech Act and PP2 Status for any of the prosodic patterns. A possible explanation is that 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 increase in the dominance of the PP1 break as an 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 not linguistically motivated,

i.e., the PP1 break in Mod versions of sentences or the OBJ break in Arg versions of sentences, 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 phonological length reasons. It is also likely that some readers are simply producing a break after each phrase as a strategy for dealing with difficult to comprehend sentences.

Speech act is a significant predictor of PP1 break dominance (β=0.31, std. error = 0.15, $p < 0.05$, see Section 3.2.4.2)., but not of any of the other prosodic outcomes jjust discussed. It is 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 no observed interaction between Speech Act and PP2 Status is discouraging for the hypothesis that what makes the Arg cases seem easier in the interrogative cases compared to the declarative is the prosody of questions.

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 provide an estimate 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, which can be taken as an indicator of processing load on the first reading.

Since 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 (n = 2) and above 25.0s (n = 5) were assumed to be implausible and omitted from the analyses reported below. Experimental data were then Winsorized by participant to bring data

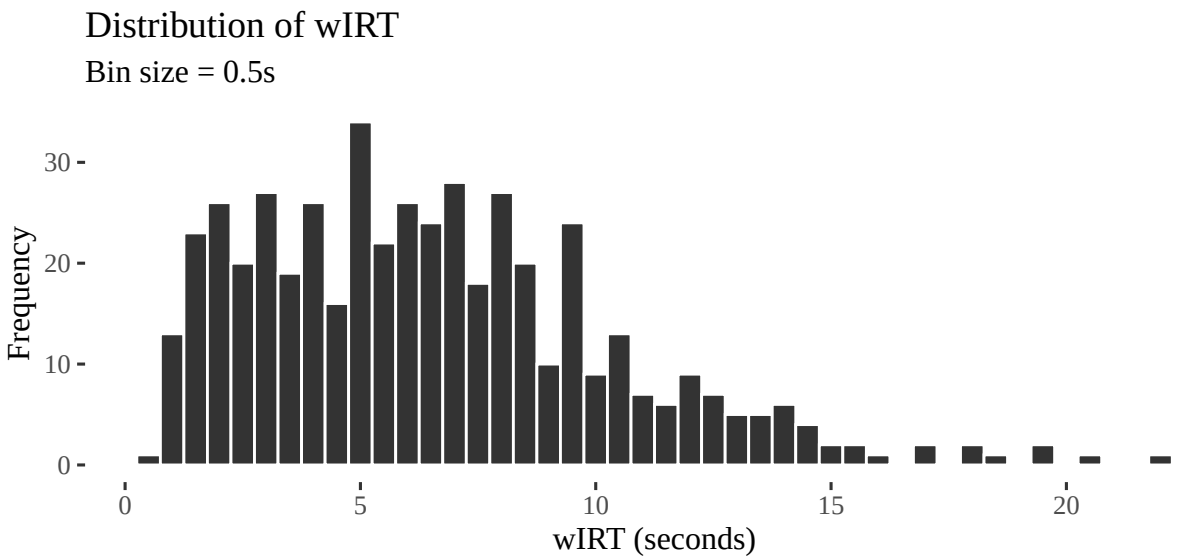

Figure 3.5: Distribution of wIRT.

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 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.

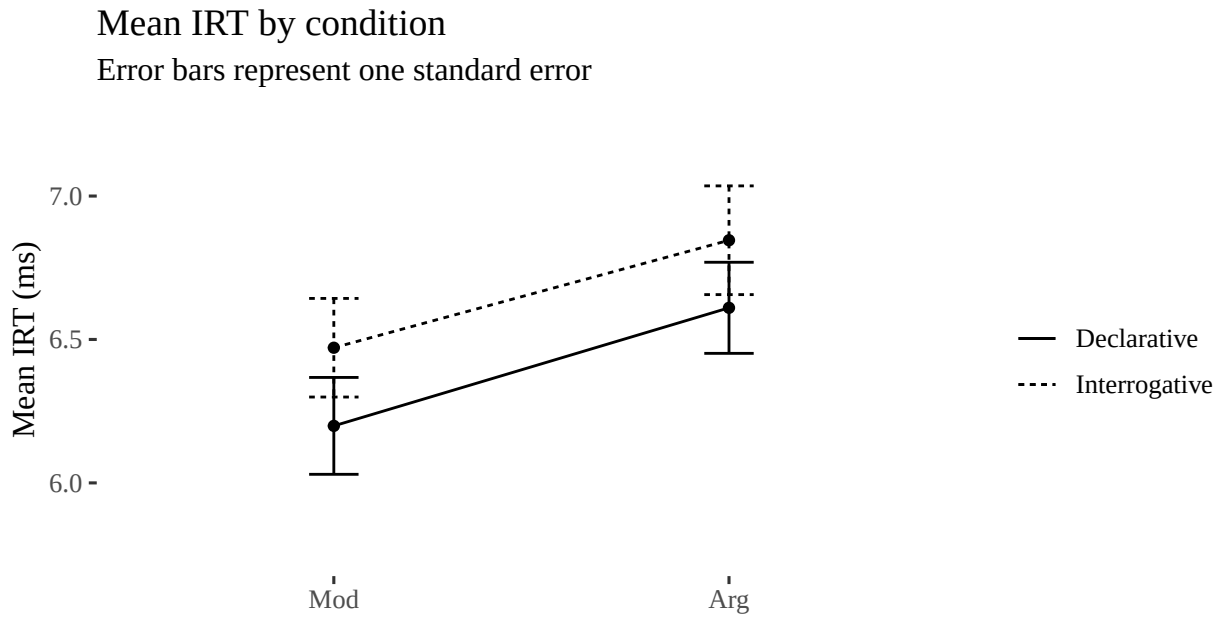

Figure 3.6: Mean IRT as a function of sentence type.

The full model for predicting IRT is one with fixed effects of Speech Act and PP2 Status and the interaction between them. This model is shown in table 3.14.

Table 3.14: 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.

The difference between the reduced and full model was not statistically significant (AIC_{FULL}=9103.5, $AIC_{REDUCED}$=9101.6, $\chi^2(1)=0.04$, $p > 0.8$). These results are not supportive of *Hypothesis 4* formalized in Section 1.6, as statistical significance was expected for the interaction of Speech Act with PP2 Status if interrogative context is indeed ameliorating for PP-attachment garden paths,

Table 3.15: 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

compared to declarative context for the same construction. Discussion of that is returned to in Section 3.4.4 below.

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 attempts to predict IRT by the PP2 head (e.g., *into* vs. *from*) found that while *into* and *onto* do not behave differently from each other, *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 that 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 3.16: 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 *of* 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 ($p > 0.7$).

3.4.4 Discussion of IRT results

It is clear from the above that the sentences containing argument PP2s tested here (ones headed by *into*/*onto*) result in longer wIRT measures than those containing Mod PP2s ($\beta = 0.46, p < 0.05$, see Section 3.4.2). It also appears that interrogativity increases wIRT regardless of the PP2 Status ($\beta = 0.33, p = 0.12$), because although that finding does not reach significance, Speech Act is included in the model with the lowest AIC. 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 (55) has only one reading, (56) has another possible reading, albeit somewhat implausible: i.e., we could imagine that in (56) *from her brother-in-law* modifies *the cookies*, while in (55) *of the minivan* cannot modify *the bicycle*.

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

(56) 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, may have 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 per Frazier & Clifton (1996) (see Section 1.2 above for a discussion of how this relates to the matter at hand), 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 differently, 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: (57) is ungrammatical, but (58) is fine; the same holds for (59) where *to Fred* is an argument, compared to (60) where *for Fred* is a modifier.

(57) $*$ I saw the student of physics and the one of chemistry arguing with each other.

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

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

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

There might also be a simpler explanation: because *of* is only two characters long, 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 sentences with four-character PP2 heads (most notably the *into*/*onto* cases), which would eliminate some of the needed study time.

Ultimately, *hypothesis 4* as originally formalized in Section 1.6 is not supported.

(61) *Hypothesis 4* The inter-reading time (IRT) will be longer for Arg sentences that are declarative than for:

a. Arg sentences that are interrogative

b. Mod sentence that are interrogative or declarative

The interaction between Speech Act and PP2 Status is not a statistically significant predictor of IRT. This also blocks a satisfying answer to whether or not the 2016 Intuition of Peckenpaugh (2016) can extend to the items tested here, since neither study found significant results for the investigated behavioral correlates of that intuition.

3.4.5 The processing cost of interrogativity

It is worth noting 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 were 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 outside the construction of interest. 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 of the sentence to mirror the experimental items: a model predicting wIRT by the presence of those PPs found minimal effect on wIRT ($\beta=0.01$, std. error = 0.22, $p = 0.96$). This indicates that the presence of a string of PPs does not independently affect the apparent processing cost of interrogativity.

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 the experimental study just reported 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:

(62) *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 had crammed the newspapers under the sofa in the trashcan*, vs., *Had Jed crammed the newspapers under the sofa in the trashcan?*)

(63) *The Current Hypothesis:* The 2016 Intuition may be extensible to PP attachment ambiguities that are syntactically disambiguated in addition to those that are pragmatically disambiguated (e.g., *He had planned to cram the paperwork in the drawer into his briefcase*,

vs., *Had he planned to cram the paperwork in the drawer into his briefcase?*).

The goal of this study was to establish whether there is evidence for (63), and to explore the implications of those findings for possible explanations of (62).

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

Ultimately, no evidence has been found to support the Current Hypothesis, that the 2016 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 2016 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 considers the evidence for and against some of the possible explanations for the 2016 Intuition reported in Peckenpaugh (2016). While evidence of a behavioral correlate has not been found, it's nonetheless interesting to consider the source of the intuition itself.

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 a failed 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 would mean that if the sentences were more prosodically similar across the Arg vs. Mod PP2 Status in the interrogative than in the declarative, the intuited reduction in difficulty of reanalysis would naturally follow. This assumes, of course, that the 2016 Intuition is eventually confirmed, despite the inconclusive results of Peckenpaugh (2016).

The findings of the current study show that there is no one-to-one mapping between prosodic structure and the four sentence types tested in this study (Q Arg, D Arg, Q Mod, D Mod). Rather, for each Sentence Type, gradient differences in occurrence of each pattern for a given sentence type were observed. 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).

(64)

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

Table 4.1 shows the distribution of the possible 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.

Table 4.1: Percent occurrence of break patterns in Reading 2 as a function of sentence type.

	Mod		Arg	
	D	Q	D	Q
OBJ break only	31.1%	31.4%	0.8%	2.5%
Both breaks	54.1%	43.0%	72.1%	71.7%
PP1 break only	14.8%	25.6%	27.0%	25.8%

While there is a larger drop in the number of utterances with both breaks from Arg to Mod for questions (71.7% - 43.0%) than for declaratives (72.1% - 54.1%), the opposite is true for the pattern

with a PP1 break by itself (25.8% - 25.6%, vs., 27.0% - 14.8%). There is an argument to be made that the OBJ breakin the Arg versions is not a true prosodic break but merely a hesitation due to a moment of confusion or for length reasons. Note that the length of the material for the current study was similar across items and may have influenced the optimal prosodic phrasing and therefore the presence or absence of some of the breaks discussed here. For explanations of length effects on syntactic parsing see, e.g., Clifton, Carlson, & Frazier (2006), Webman-Shafran & Fodor (2016), and Dinçtopal Deniz & Fodor (2017).

If the prosodic structures in (65), (66), (67) and (68) (see below) are assumed to be ideal, then an explanation for the 2016 intuition (and by extension, the same explanation for syntactically disambiguated cases) presents itself. The symbol "%" is being used in these examples to represent a prominent prosodic break.

(65) *D Arg*: He had crammed [$_{OBJ}$ the newspapers] [$_{PP1}$ under the sofa] % [$_{PP2}$ into the trashcan].

(66) *D Mod*: He had crammed [$_{OBJ}$ the newspapers] % [$_{PP1}$ under the sofa [$_{PP2}$ in the guestroom]].

The two declarative versions differ from each other in that the Arg case is ideally pronounced with a major break after *under the sofa* to mark the argument attachment of PP2; whereas the Mod case does not require that break in the ideal pronunciation (though it may sometimes occur for length reasons).That is, the PP1 break differentiates the two syntactic structures and signals the attachment site of PP2 in the Arg case.

For the interrogatives, though, this contrast is obscured by the need to apply a rising prosodic contour over the final nuclear accent in the sentence: i.e., PP2, resulting in not a change in syntactic branching direction, but the anticipation of a tonal change, which might sound very similar to a prosodic break. The symbol "⇑" in (67) and (68) indicates the start of the rising contour, and "⇒" the point in the string where preparation might begin.

(67) *Q Arg*: Had he crammed [$_{OBJ}$ the newspapers] [$_{PP1}$ under the sofa] % [$_{PP2}$ into the ⇑ trashcan]?

(68) *Q Mod*: Had he crammed [$_{OBJ}$ the newspapers] [$_{PP1}$ under the sofa ⇒ [$_{PP2}$ in the ⇑

guestroom]]?

The critical issue here is that in (68), the absence of a syntactically-motivated break before PP2 might be obscured by the hesitation necessitated by preparing to start the final-rising contour of a question two words later. It's not impossible that the need to make the mechanical preparation necessary to execute a rising contour could result in a hesitation or pause at the preceding syntactic juncture most welcoming to it, the PP1 break position. While in (68) PP2 is a sub-constituent of PP1, and so the final phrase of the VP is actually PP1, the final nuclear accent in (68) still falls on the NP within PP2. Thus the final high rise is anchored to the same phrase in both interrogative versions (Arg and Mod).

If the PP1 break is treated as the main indicator of the prosodic structure, then the smaller difference across PP2 Status for interrogatives (0.2%) than for declaratives (12.2%) does, though perhaps weakly, leave open the door for the Bader (1998) style explanation of the hunch that interrogative PP-attachment garden paths are easier to comprehend than declarative ones. Ultimately, though, logistic regression models showed no significant effect of the interaction between Speech Act and PP2 Status, so it is safest to assume that this difference is just noise, and that the explanation for the intuition is not prosodic.

4.2.2 A processing illusion

If the above explanation of a null result proves not to be viable in subsequent study, another possibility to consider is that the intuition itself is illusory in nature, i.e., it is not the case that interrogative PP-attachment garden paths (either pragmatically or syntactically disambiguated) are easier to parse than declarative ones, but only appear to be so. . Because the added complexity of the interrogative version is distracting the reader with a different sort of difficulty. That is, it may be that readers are being distracted by the processing cost of interrogativity, and thereby are less sensitive to the cost of reanalysis. While the parser is struggling with two additive increases to complexity as compared to the D Mod sentences (i.e., Arg PP2 Status and interrogative status), the

reader is aware only of the more immediately obvious difference, that of interrogativity, and fails to notice the structural reanalysis that the parser is undertaking.

4.2.3 A semantic/pragmatic explanation

As discussed in Section 1.3, there are substantial pragmatic and semantic differences between the interrogative and declarative versions of the construction the current study addresses. It is possible that those difference, rather than or in addition to the prosodic differences, are what lead to the 2016 Intuition. The differing properties of focus between questions and declaratives, for exampole, may lead the grammar to generate a set of alternatives for some portion of the sentence, i.e., if the question is, "Had she intended to put the bicycle on the roof rack into the garage," and that question is interpreted to mean, "where did she put the bicycle?" a list of alternatives to *into the garage* will have to be created, or if the parse has not been correctly completed, alternatives to *on the roof rack into the minivan*. It may be that the ungrammatical nature of *on the roof rack into the minivan* can be ignored if only a list of alternatives is needed, or that generating the list of alternatives in some way obscures the difficulty of processing the sentence. This is not the only semantically or pragmatically based explanation available; it suffices to say that it is possible for semantics and pragmatics to play a role in explaining why there would is a 2016 Intuition at all.

4.3 Conclusions and future directions

In this section, I will discuss some of the potential issues raised by the current study, and what I see as fruitful avenues for future research.

4.3.1 Inter-item reading time (IRT) findings

A number of interesting findings are supported by IRT and the Double Reading paradigm used in the current study.

While IRT did not detect a behavioral correlate of the motivating hunch for this study, it did show a

robust effect of PP2 Status (Arg vs. Mod), indicating that it may yet be a useful diagnostic for parsing difficulty in general.

The IRT data reported also support prior findings that interrogatives are generally slower to process or more difficult to comprehend than related declarative sentences, as noted by Mehler (1963) and Peckenpaugh (2016) (see Section 3.4.5 above). .

The fact that reading (Reading 1 vs. Reading 2) frequently had an effect on prosodic phrasing also supports the finding of Fodor et al. (2019) that ”cold” and previewed readings have different properties and can provide different sorts of psycho-linguistic evidence; thus item preview should be tightly controlled and documented.

4.3.2 Hesitations vs. prosodic breaks

An issue with the findings of this study is the difficulty in differentiating between linguistically-motivated prosodic breaks and hesitations. As shown by the reliability data reported in Section 2.8.1, linguistically trained judges will not necessarily be able to agree where breaks fall and which breaks are stronger, perhaps in part because of the difficulty in differentiating between prosodic breaks and hesitations.

It can be difficult to capture a prosodic break in physical terms (see, e.g., Ladd (2008)), and there is no guarantee that examining the wave forms instrumentally would fare any better than judges' intuitions in distinguishing between hesitation and prosodic breaks. That said, it may be possible to distinguish the two sorts of breaks: for example, it might be that boundary tones or segmental lengthening are typically not part of a hesitation break, but are a part of a prosodic break, for example.

In 4.3.3.2 I propose that an event-related potential (ERP) paradigm might hold the key for distinguishing between the two.

4.3.3 Future work

Some future studies that could further this line of research are outlined below.

4.3.3.1 Embedded questions

The explanation for the intuition that PP-attachment garden paths are less difficult to parse in interrogatives than in declaratives (if indeed it is correct) must either be prosodic, or else semantic/pragmatic. It seems very unlikely that the minor syntactic difference across the two sentence types (i.e., subject-auxiliary inversion) could be the reason for a difference in perceived ease of understanding. More plausibly, it is either the difference in the intonational contour between the two sentences types, or else it is some aspect of the meaning or meta-linguistic difference between interrogatives and declaratives. The study reported here has shown evidence that the intuition cannot easily be explained entirely by the prosodic differences between questions and declaratives; that possibility remains weak at best. It seems that the most important next step in explaining the aforementioned intuition would be to look at the same phenomenon in embedded questions vs. embedded declarative clauses, where prosody would not be at play but the semantic/pragmatic differences should remain. For example:

(69) *Em Q Arg:* He asked her if she had decided to cram the old newspapers under the couch in the wastebasket.

(70) *Em Q Mod:* He asked her if she had decided to cram the old newspapers under the couch in the guestroom.

(71) *Em D Arg:* She told him that she had decided to cram the old newspapers under the couch in the wastebasket.

(72) *Em D Mod:* She told him that she had decided to cram the old newspapers under the couch in the guestroom.

The prosody of an embedded question, as in (69), does not differ from that of a sentence like (71);

but, the semantic properties and some of the pragmatic properties of the embedded clause in (69) are the same as in the Q Arg sentences from the study just reported.

4.3.3.2 Event-related potentials

An event-related potential study of the phenomenon could provide a number of useful insights. It has been shown that the closure-positive shift (CPS), an ERP component that is elicited by the presence of a prosodic break or a comma, is present even in silent reading (Drury, Baum, Valeriote, & Steinhauer, 2016). Together with the well known P600 (a positive amplitude shift approximately 600ms after the anamoly), associated with syntactic anomalies (Neville, Nicol, Barss, Forster, & Garrett, 1991) and N400 (a negative amplitude shift approximately 400ms after the anamoly), associated with semantic anomalies (Kutas, Van Petten, & Kluender, 2006) create a set of measures that have a number of important implications for research on the phenomena at hand.

Specifically, it might be that these ERP components can be used to:

A. Distinguish between hesitation and linguistically motivated prosodic breaks.

B. If (A) proves fruitful, provide reference data with which to train either linguists or computer models to better distinguish between the two types of breaks.

C. Determine whether the disambiguation style of Peckenpaugh (2016) compared to the current study differs in terms of being syntactic or semantic.

To accomplish (A), one could time lock an ERP measure to the PP1 break position for the voice of a participant who is reading sentence like this for the current study aloud, and examine the resulting potentials. If a CPS is detected, that might be taken to mean that the participant has determined that a break there is linguistically motivated, and the recording of their voice might be scrutinized as an example of a true prosodic break. If an N400 or P600 is detected, that might mean that any break at that position should be considered a hesitation rather than a linguistically motivated break. It's also possible for these components to co-occur: Steinhauer (2003) found that the CPS and P600 components can be additive. In that case, it might be that the participant has both been

garden pathed and also felt that a break was linguistically motivated. The recordings created in the pursuit of (A), if they could be plausibly thought to represent the two categories of prosodic breaks, could then be studied and/or used to train either linguists or computer models to better distinguish hesitation from prosidic breaks. See Rosenberg (2009) for discussion of the use of computer models to categorize prosodic events.

Finally, (C) would follow much the same methodology, and would seek to find a difference between, e.g., *in* PP2 disambiguation compared to *into* disambiguating PP2s. If N400 components were found to occur at the PP1 break position for Peckenpaugh (2016) type sentences but P600 were found to occur there for the types of sentences tested in the current study, that finding would support the idea that the two are disambiguated in a different fashion. Other patterns are also possible, and it might be found that some items in each study are disambiguated in a way that triggers an N400 component while others trigger a P600 component. Whatever the findings, they could prove useful for refining the items used in the future of the PP-attachment research paradigm.

4.3.4 Summary

In sum, the current study has not found evidence that supports the extension of the 2016 Intuition that interrogative garden paths are easier to process than declarative garden paths to syntactically disambiguated items. This is similar to the null findings of Peckenpaugh (2016). Evidence has been found of an effect on Inter-reading time (IRT) of the Mod vs. Arg attachment of PP2 in both declarative and interrogative cases. Also established here is an effect of Speech Act on the IRT for filler items. There are systematic differences in the way that the different sentence types are prosodically structured, but the correspondence between sentence type and prosodic pattern is gradient rather than categorical. It is possible that a Bader-type account can explain the intuition investigation, or that it is illusory, or that the intuition is due to semantic or pragmatic factors, or to any combination of such factors. Further study via behavioral, eye-tracking, or ERP methodology are likely to be fruitful in further understanding the phenomenon.

Appendix A

Experimental items

Table A.1: Experimental items in four versions

Version	Text
Q Arg	Had she decided to cram the cookies in the basket into her jacket pocket?
D Arg	She had decided to cram the cookies in the basket into her jacket pocket.
Q Mod	Had she decided to cram the cookies in the basket from her brother-in-law?
D Mod	She had decided to cram the cookies in the basket from her brother-in-law.
Q Arg	Had she decided to put the child on the rocking horse onto the see-saw?
D Arg	She had decided to put the child on the rocking horse onto the see-saw.
Q Mod	Had she decided to put the child on the rocking horse from his parents?
D Mod	She had decided to put the child on the rocking horse from his parents.
Q Arg	Had he decided to set the board games on the floor onto the card table?
D Arg	He had decided to set the board games on the floor onto the card table.
Q Mod	Had he decided to set the board games on the floor of the living room?
D Mod	He had decided to set the board games on the floor of the living room.
Q Arg	Had he decided to stick the large check in the envelope into her wallet?
D Arg	He had decided to stick the large check in the envelope into her wallet.
Q Mod	Had he decided to stick the large check in the envelope from her church?
D Mod	He had decided to stick the large check in the envelope from her church.
Q Arg	Had he intended to cram the paperwork in the drawer into his boss's desk?
D Arg	He had intended to cram the paperwork in the drawer into his boss's desk.
Q Mod	Had he intended to cram the paperwork in the drawer of his filing cabinet?
D Mod	He had intended to cram the paperwork in the drawer of his filing cabinet.
Q Arg	Had he intended to put the bicycle on the roof rack into the garage?
D Arg	He had intended to put the bicycle on the roof rack into the garage.
Q Mod	Had he intended to put the bicycle on the roof rack of the minivan?
D Mod	He had intended to put the bicycle on the roof rack of the minivan.

Table A.1: Experimental items in four versions *(continued)*

Version	Text
Q Arg	Had she intended to set the clothes in the hamper onto the dresser?
D Arg	She had intended to set the clothes in the hamper onto the dresser.
Q Mod	Had she intended to set the clothes in the hamper from his sister?
D Mod	She had intended to set the clothes in the hamper from his sister.
Q Arg	Had she intended to stick the letter in the mailbox onto the proper stack?
D Arg	She had intended to stick the letter in the mailbox onto the proper stack.
Q Mod	Had she intended to stick the letter in the mailbox of the vice president?
D Mod	She had intended to stick the letter in the mailbox of the vice president.
Q Arg	Had she planned to cram the stolen files in the wall-safe into a suitcase?
D Arg	She had planned to cram the stolen files in the wall-safe into a suitcase.
Q Mod	Had she planned to cram the stolen files in the wall-safe of their hideout?
D Mod	She had planned to cram the stolen files in the wall-safe of their hideout.
Q Arg	Had she planned to put the jelly beans in the window onto a fancy dish?
D Arg	She had planned to put the jelly beans in the window onto a fancy dish.
Q Mod	Had she planned to put the jelly beans in the window of his candy store?
D Mod	She had planned to put the jelly beans in the window of his candy store.
Q Arg	Had he planned to set the appetizers on the platter onto the buffet?
D Arg	He had planned to set the appetizers on the platter onto the buffet.
Q Mod	Had he planned to set the appetizers on the platter from his cousin?
D Mod	He had planned to set the appetizers on the platter from his cousin.
Q Arg	Had he planned to stick the post-it note on the handout onto his notebook?
D Arg	He had planned to stick the post-it note on the handout onto his notebook.
Q Mod	Had he planned to stick the post-it note on the handout from the lecture?
D Mod	He had planned to stick the post-it note on the handout from the lecture.
Q Arg	Had he wanted to cram the newspapers under the sofa into the wastebasket?
D Arg	He had wanted to cram the newspapers under the sofa into the wastebasket.
Q Mod	Had he wanted to cram the newspapers under the sofa from the thrift store?
D Mod	He had wanted to cram the newspapers under the sofa from the thrift store.
Q Arg	Had he wanted to put the photo on the coffee table onto the mantelpiece?
D Arg	He had wanted to put the photo on the coffee table onto the mantelpiece.
Q Mod	Had he wanted to put the photo on the coffee table from his grandfather?
D Mod	He had wanted to put the photo on the coffee table from his grandfather.
Q Arg	Had she wanted to set the textbooks on the top shelf into the file box?
D Arg	She had wanted to set the textbooks on the top shelf into the file box.
Q Mod	Had she wanted to set the textbooks on the top shelf of the book shelf?
D Mod	She had wanted to set the textbooks on the top shelf of the book shelf.
Q Arg	Had she wanted to stick the golf clubs in the back room into the closet?

Table A.1: Experimental items in four versions *(continued)*

Version	Text
D Arg	She had wanted to stick the golf clubs in the back room into the closet.
Q Mod	Had she wanted to stick the golf clubs in the back room of their condo?
D Mod	She had wanted to stick the golf clubs in the back room of their condo.

Appendix B

Filler items

Table B.1: Filler items with trailing PPs

Version	Text
+PP D	He had intended to enter the expenses from the trip into a spreadsheet.
+PP Q	Had he intended to enter the expenses from the trip into a spreadsheet?
-PP D	He had intended to do the work that the boss asked a coworker to do.
-PP Q	Had he intended to do the work that the boss asked a coworker to do?
+PP D	He had intended to sell his collection of baseball cards from his childhood.
+PP Q	Had he intended to sell his collection of baseball cards from his childhood?
-PP D	He had intended to replace the crackers he ate while he was house sitting.
-PP Q	Had he intended to replace the crackers he ate while he was house sitting?
+PP D	She had planned to tell the student in private about his failing grade.
+PP Q	Had she planned to tell the student in private about his failing grade?
-PP D	She had planned to finish preparing dinner while the guests were chatting.
-PP Q	Had she planned to finish preparing dinner while the guests were chatting?
+PP D	She had planned to pack a ham sandwich on rye bread into her lunchbox.
+PP Q	Had she planned to pack a ham sandwich on rye bread into her lunchbox?
-PP D	She had planned to build herself a new computer when she got her paycheck.
-PP Q	Had she planned to build herself a new computer when she got her paycheck?
+PP D	She had wanted to complete the race for charity in record time.
+PP Q	Had she wanted to complete the race for charity in record time?
-PP D	She had wanted to bring her son when she attended the next conference.
-PP Q	Had she wanted to bring her son when she attended the next conference?
+PP D	She had wanted to find a rare butterfly on their hike in the rainforest.
+PP Q	Had she wanted to find a rare butterfly on their hike in the rainforest?

Table B.1: Filler items with trailing PPs *(continued)*

Version	Text
-PP D	She had wanted to tell her friends that she was selling her vacation home.
-PP Q	Had she wanted to tell her friends that she was selling her vacation home?
+PP D	She had decided to break the class into teams of six students each.
+PP Q	Had she decided to break the class into teams of six students each?
-PP D	He had decided to do the needed repairs on the broken-down van himself.
-PP Q	Had he decided to do the needed repairs on the broken-down van himself?
+PP D	She had decided to instruct the staff on proper etiquette for formal dining.
+PP Q	Had she decided to instruct the staff on proper etiquette for formal dining?
-PP D	He had decided to advise that his newest patient seek a second opinion.
-PP Q	Had he decided to advise that his newest patient seek a second opinion?
+PP D	He had forgotten to try the famous pastry in the restaurant of the fancy hotel.
+PP Q	Had he forgotten to try the famous pastry in the restaurant of the fancy hotel?
-PP D	She had forgotten to report that the clerk was ignoring her request.
-PP Q	Had she forgotten to report that the clerk was ignoring her request?
+PP D	He had forgotten to tack the pamphlet on hygiene onto the notice board.
+PP Q	Had he forgotten to tack the pamphlet on hygiene onto the notice board?
-PP D	She had forgotten to lock the gate that was supposed to be kept closed.
-PP Q	Had she forgotten to lock the gate that was supposed to be kept closed?
+PP D	She had meant to arrange the files in alphabetical order for her boss.
+PP Q	Had she meant to arrange the files in alphabetical order for her boss?
-PP D	She had meant to write up the performance reviews to give her employees.
-PP Q	Had she meant to write up the performance reviews to give her employees?
+PP D	She had meant to place a suggestion box onto the front desk of the clinic.
+PP Q	Had she meant to place a suggestion box onto the front desk of the clinic?
-PP D	She had meant to try to get the program to run on the new operating system.
-PP Q	Had she meant to try to get the program to run on the new operating system?
+PP D	He had needed to request some money from his father-in-law for the remodel.
+PP Q	Had he needed to request some money from his father-in-law for the remodel?
-PP D	He had needed to upgrade his ticket when he changed his travel plan.
-PP Q	Had he needed to upgrade his ticket when he changed his travel plan?
+PP D	He had needed to set the vegan cookies onto serving trays for the party.
+PP Q	Had he needed to set the vegan cookies onto serving trays for the party?
-PP D	He had needed to beg to get his old job back when his investment failed.
-PP Q	Had he needed to beg to get his old job back when his investment failed?

Table B.1: Filler items with trailing PPs *(continued)*

Version	Text
+PP D	He had remembered to add the section into the handboook for the meeting.
+PP Q	Had he remembered to add the section into the handboook for the meeting?
-PP D	He had remembered to tell the office manager to order more coffee filters.
-PP Q	Had he remembered to tell the office manager to order more coffee filters?
+PP D	He had remembered to move the gifts from the baby shower onto the bed.
+PP Q	Had he remembered to move the gifts from the baby shower onto the bed?
-PP D	He had remembered to circulate the latest job posting his company had sent.
-PP Q	Had he remembered to circulate the latest job posting his company had sent?

Table B.2: Filler items with no trailing PPs

	Version	Text
3	TRUE	-PP D
4	TRUE	-PP Q
7	TRUE	-PP D
8	TRUE	-PP Q
11	TRUE	-PP D
12	TRUE	-PP Q
15	TRUE	-PP D
16	TRUE	-PP Q
19	TRUE	-PP D
20	TRUE	-PP Q
23	TRUE	-PP D
24	TRUE	-PP Q
27	FALSE	-PP D
28	FALSE	-PP Q
31	FALSE	-PP D
32	FALSE	-PP Q
35	FALSE	-PP D
36	FALSE	-PP Q
39	FALSE	-PP D
40	FALSE	-PP Q
43	FALSE	-PP D
44	FALSE	-PP Q
47	FALSE	-PP D
48	FALSE	-PP Q
51	FALSE	-PP D
52	FALSE	-PP Q
55	FALSE	-PP D
56	FALSE	-PP Q
59	FALSE	-PP D
60	FALSE	-PP Q
63	FALSE	-PP D
64	FALSE	-PP Q

Appendix C

Recruitment notice

You will be asked about your reading habits and then asked to read complex sentences out loud while being audio recorded. Recordings of your voice will be analyzed, but will be kept strictly confidential. The process will take no more than 1 hour. Note that the study takes place in Queens Hall, which is about half a mile from the main Queens campus. See directions on the QC website, URL below, for how to get here. The room is 335D, on the third floor. Entrance to the building is in the back.

Appendix D

Instructions to participants

Thank you kindly for your participation. In this study, you are being asked to read complex sentences out loud, twice each. It is very important that you follow these guidelines for each of your readings.

First reading: Begin reading immediately, without giving yourself a chance to look ahead. Imagine you are a television reporter reading an urgent update from a teleprompter. You must be as quick as possible, without taking any time to read ahead. You want to sound natural if you can, but it is more important to not delay. These sentences are complicated and potentially confusing. It's very important that you read the sentence out loud as soon as it appears. It's OK if you make mistakes or don't understand, that is an important part of what I want to know. Do the best you can, and remember you have another chance to read it.

Second reading: This time you have the luxury of pacing yourself as you please. Imagine you are providing a voice-over for a documentary. You want to sound conversational and clear, without being overly dramatic or formal. Study the sentence as long as you like, and be sure that you understand it before you begin reading. It is most important to sound natural, without worrying about how long it takes to prepare.

The experiment will begin with brief instructions, recapping what you are reading now. There will then be a practice session to get you comfortable with the task and a chance for you to ask any questions you have. Finally, after your questions are answered, the study will begin in earnest.

Each sentence will follow the same pattern. You will be presented with a screen which displays a series of plus signs. This indicates that the system is ready and that you should press the button labeled "START" when you are ready to read a sentence. As soon as you press the button, the sentence will appear and you should begin your first reading. After you have completed the reading, press the button labeled "NEXT." You should allow a small amount of time after you finish and before you hit "NEXT," to ensure that the recording is not cut off too early.

Once you have pressed "NEXT," you will see a brief instructions slide to help you keep track of where you are. You should then press "START" and begin preparing to read the second time. The background color will change to confirm that the computer has registered your key press. Once you're ready, read the sentence aloud for the second time and then press "DONE." Once again, be

sure not to cut yourself off. Wait a moment after you finish reading before pressing “DONE.”

You are not being judged or measured in any way. Rather, we are interested in how these sentences are pronounced by native speakers of English. Any confusion you have or mistakes you make are interesting properties of the sentences, not failings of you, the speaker.

The keys used during the experiment are clearly labeled, but the function of each key is listed below for your reference. There is no hurry for pressing the keys. The only timing of importance is that you begin reading as quickly as possible after pressing “START.” The task should take no longer than one hour

Table D.1: Table of keyboard mappings

Label	Position	Description
Start	Left shift	Revewal a sentence and begin your reading.
Next	Right shift	End your first reading.
Done	Thumb pad	End your second reading and prepare for the next sentence.

Appendix E

Instructions for prosodic coding

Before you begin describing the recordings for a given speaker, please familiarize yourself with that speaker. To do so, please listen to recordings numbered 46-48 and 24-27.

Next, move on to describings the recordings numbered 1-16. Please listen to them in the following pattern: begin with either 1Y or 1X, and listen to the recordings sequentially (or reverse sequentially), and alternate between X and Y versions. Then, repeat the process for the inverse versions (X vs. Y). Please then listen to the next speaker, beginning with 16X or Y, and then listen in reverse sequence, alternating X vs. Y, and then again repeat for the other half. In this way, please alternate across speakers between listening to X or Y first as well as 1 or 16 first.

For each recording, please respond in the spreadsheet using the following guidelines for the columns. Each recording should get its own row.

- Speaker ID: This should be the name of the directory in which the recording exists.
- Recording ID: This should be the filename of the recording being described
- X or Y: This should be the last character of the filename, either X or Y.
- First recording for speaker: This should indicate which recording you started with first, which will allow me to deduce the pattern you used to listened to the recordings, per above, e.g. 1X, 1Y, 16X or 16Y.

For columns E-K, Consider a given sentence to be divided into regions, as in the following example:

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

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.

- Break after V?: Please indicate whether or not you think there is a prosodic boundary after the verb cluster(at the right edge of the last/main verb).

- Break after OBJ?: Please indicate whether or not you think there is a prosodic boundary after the first NP in the object region (at the right edge of the first NP in the object region).
- Break after PP1?: Please indicate whether or not you think there is a prosodic boundary after the first NP in the PP1 region (at the right edge of the first NP in the PP region).
- Strongest break? Please indicate which of the breaks (columns E-G where you indicated YES) you think is strongest. If two breaks are of equal strength and are stronger than a third, indicate NONE as strongest. If two breaks are of equal strength and are weaker than a third, indicate that third break as strongest. If all breaks are the same strength, indicate NONE as strongest.
- Weakest break? Please indicate which of the breaks (columns E-G where you indicated YES) you think is weakest. If two breaks are of equal strength and are weaker than a third, indicate NONE as weakest. If two breaks are of equal strength and are stronger than a third, indicate that third break as weakest. If all breaks are the same strength, indicate NONE as weakest
- Struggle?: Indicate whether or not the speaker appears to have had difficulty reading the sentence. This should be relative to their baseline reading fluency, so if a person is hesitant every time, hesitance should not be enough to indicate a struggle.
- Start of struggle: indicate the region in which you first notice the speaker struggling. *Question?: indicate simply whether or not the recording sounds like a question, prosodically (e.g., final rise is present).

References

Ashby, J., Yang, J., Evans, K. H., & Rayner, K. (2012). Eye movements and the perceptual span in silent and oral reading. *Attention, Perception, and Psychophysics*, *74*(4), 634–640.

Bader, M. (1998). Prosodic influences on reading syntactically ambiguous sentences. In *Reanalysis in sentence processing* (pp. 1–46). Springer.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2019). *Lme4: Linear mixed-effects models using 'eigen' and s4.* Retrieved from https://CRAN.R-project.org/package=lme4

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, *3*, 30.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the Development of Language*, *279*(362), 1–61.

Chomsky, N. (2014). *The minimalist program*. MIT press.

Clifton, C., Jr. (1988). Restrictions on late closure: Appearance and reality. *6th australian language and speech conference*, 19–21.

Clifton, C., Jr., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, *30*(2), 251–271.

Clifton, C., Carlson, K., & Frazier, L. (2006). Tracking the what and why of speakers' choices: Prosodic boundaries and the length of constituents. *Psychonomic Bulletin & Review*, *13*(5), 854–861.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in spanish. *Cognition*, *30*(1), 73–105.

Den Dikken, M. (2006). *Relators and linkers: The syntax of predication, predicate inversion, and copulas* (Vol. 47). MIT press.

Dinçtopal Deniz, N., & Fodor, J. D. (2017). Phrase lengths and the perceived informativeness of prosodic cues in turkish. *Language and Speech*, *60*(4), 505–529.

Drury, J. E., Baum, S. R., Valeriote, H., & Steinhauer, K. (2016). Punctuation and implicit prosody in silent reading: An erp study investigating english garden-path sentences. *Frontiers in Psychology*, *7*, 1375.

Falk, T. H., & Chan, W.-Y. (2006). Nonintrusive speech quality estimation using gaussian mixture models. *IEEE Signal Processing Letters*, *13*(2), 108–111.

Fiengo, R. (2007). *Asking questions: Using meaningful structures to imply ignorance*. Oxford University Press.

Fodor, J. D. (2002). Psycholinguistics cannot escape prosody. *Speech prosody 2002, international conference*.

Fodor, J. D., Macaulay, B., Ronkos, D., Callahan, T., & Peckenpaugh, T. (2019). Center-embedded sentences: An online problem or deeper? In *Grammatical approaches to language processing* (pp. 11–28). Springer.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*.

Frazier, L., & Clifton, C., Jr. (1996). *Construal*. MIT Press.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, *4*(4), 232–237.

Hedberg, N., Sosa, J. M., & Görgülü, E. (2017). The meaning of intonation in yes-no questions in american english: A corpus study.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of american english. *The Journal of the Acoustical Society of America*, *128*(2), 839–850.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*(1), 15–47.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified ii (1994–2005). In *Handbook of psycholinguistics* (pp. 659–724). Elsevier.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2019). *LmerTest: Tests in linear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=lmerTest

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Laubrock, J., & Kliegl, R. (2015). The eye-voice span during reading aloud. *Frontiers in Psychology*, *6*(1432). https://doi.org/10.3389/fpsyg.2015.01432

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Maynell, L. A. (1999). Effect of pitch accent placement on resolving relative clause ambiguity in english. *Poster presented at the 12th annual cuny conference on human sentence processing, new york*, 18–20.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*(6), 578–586.

Mehler, J. (1963). Some effects of grammatical transformations on the recall of english sentences. *Journal of Verbal Learning and Verbal Behavior*, *2*(4), 346–351.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of*

Cognitive Neuroscience, *3*(2), 151–165.

Peckenpaugh, T. (2016). *Interrogative context and PP-attachment ambiguities*.

Prince, A., & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 358–374.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C., Jr. (2012). *Psychology of reading*. Psychology Press.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events*. Columbia University.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169–182.

Selkirk, E. O. (1986). *Phonology and syntax: The relation between sound and structure*. MIT Press (MA).

Selkirk, E. O. (2011). The syntax-phonology interface. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The handbook of phonological theory* (Vol. 2, pp. 435–483). Oxford: Blackwell Publishing.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language*, *86*(1), 142–164.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America*, *64*(6), 1582–1592.

Truckenbrodt, H. (1999). On the relation between syntactic phrases and phonological phrases.

Wax, M., & Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(2), 387–392.

Webman-Shafran, R., & Fodor, J. D. (2016). Phrase length and prosody in on-line ambiguity resolution. *Journal of Psycholinguistic Research*, *45*(3), 447–474.