DO NOT DISTRIBUTE: PP-attachment ambiguities

Tyler J. Peckenpaugh

2019-08-04

Contents

Abstract

Write abstract here

0.1 TODO

- Abstract
- Acknowledgements
- More on the significance of IRT to the study at hand
- Hand measurement for IRT reliability
- First citation *per chapter* should include all authors (rather than et al)

WORK IN PROGRESS. DO NOT DISTRIBUTE.

Chapter 1

Introduction

Before the details of the current research can be outlined, it's first necessary to explain some of the terms and mechanisms involved. Most critically, it's necessary to be clear on what a garden path is, and what causes the phenomena.

1.1 What is a garden path?

The occurrence of garden-path (GP) effects are generally attributed to a clash between structure-first parsing strategies, e.g. Frazier (1979)'s *Minimal Attachment*, and the actual intended structure of the sentence in question.

If we assume that the parser builds structure by following these types of constraints or rules in a programmatic way, without concern for meaning, it's easy to see how one might end up with a parse that denotes an implausible situation, or even an ungrammatical sentence. 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.”

(1) 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]]

The attempted parse represented in structure (8) crashes, as it isn“t possible to incorporate the final word fell in a grammatical way. In order to obtain the grammatical parse, (8 b) where the matrix subject is the horse raced past the barn, almost the entire structure of (8 a) must be discarded. The subject is not the horse, and the main verb is not raced. Rather, the subject is a noun phrase (NP) containing a reduced relative clause raced past the barn. Only after remediation can fell be properly incorporated as the matrix verb. This finally results in a grammatical sentence, with a structure comparable to,”The horse (that was) raced past the barn was hungry.” Garden-path effects can result any time the assumptions made by the parser clash with the actual structure of a sentence.

1.2 PP-attachment garden paths

The type of garden-path sentence that this study is concerned with centers around the temporarily ambiguous attachment sites for a prepositional phrase (PP).

Consider:

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

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

The sentences in (2) and (3) contain a temporary ambiguity with regard to the attachment site of in the drawer (PP1). It can either attach high, as an argument of the verb (i.e., the paperwork is being crammed in the drawer), or else low, as a modifier of the direct object (i.e. the paperwork in the drawer is being crammed into his briefcase). Sentence (2) differs from (3) in the ultimate resolution of that ambiguity, although in both cases disambiguation is triggered by the preposition heading the second PP, PP2 (into/of). In (2), PP1 must ultimately attach high, as the locative argument of the verb, in order for a plausible parse to be obtained. This is likely to cause some difficulty for the reader or listener of the sentence, if structure-first parsing strategies like *Minimal Attachment* are obeyed. The parser will have attached PP1 as the verb"s locative argument, not knowing PP2 was to come. The resulting parse crashes, because with PP1 attached high, PP2 has nowhere to attach but as a modifier of the noun in PP1 (*[$_{\mathrm{NP}}$ the drawer into his briefcase]), or else as a modifier of the verb/sentence (*[$_{\mathrm{VP}}$ he did so into his briefcase]). The preference for VP attachment of PP1 in these kinds of sentences is supported by experimental data in studies from Rayner, Carlson, & Frazier (1983), Clifton, Speer, & Abney (1991) and Schutz and Gibson (1999).

The other sentence, (3), should not represent any difficulty. The ambiguity is resolved such that PP1 can, and in fact must, attach high, since the subsequent PP of her filing cabinet (PP2) cannot plausibly attach as a sentential modifier, nor a verbal argument or modifier, and the sentence is incomplete without a destination of the cram action. By examining these sentences, the study proposed aims to expand understanding of human sentence processing in general, and specifically the relationship between syntactic structure and prosodic structure. Further, this will explore any contrast there might be in the impact of declarative and interrogative contexts on parsing, which has been under studied.

1.3 An observation

These attachment ambiguities which are difficult to parse in the declarative, appear to be less difficult to parse when presented in yes-no interrogative context. Consider again the somewhat difficult to process sentence in (2), repeated below as (4) for convenience.

(4) He had planned to cram the paperwork [$_{\text{PP}_1}$ in the drawer] [$_{\text{PP}_2}$ into his briefcase].

The mechanics of this garden path are discussed in section 2.1. The point to consider here is that, according to the intuition of many native speakers of American English, the difficulty of (4) is at least somewhat lessened in (5):

(5) Had he planned to cram the paperwork [$_{\text{PP}_1}$ in the drawer] [$_{\text{PP}_2}$ into his

briefcase]?

That is, while both sentences contain a temporary ambiguity for the attachment of PPs occurring in string-linear sequence at the end of an utterance, it is more likely in (5) that a listener or reader will come away with a plausible interpretation. This study explores the factors involved in why that is, and seeks to uncover a behavioral correlate of this intuition. If the existence and mechanics of this effect can be pinned down, it will lend insight into what information the parser has access to when making decisions, or when repairing broken parses. This initial observation reveals a robust program of research with many interwoven questions.

One of the many questions is: can speaker behavior be shown to reflect the intuition? To say it another way, is there some behavior that is correlated with the reduction in difficulty of parsing PP-attachment garden paths that can be measured? Peckenpaugh (2016) represents one attempt to find such a correlate, and this paper continues that search.

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

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

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

Chapter 2

Background

As briefly mentioned earlier, "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-focused parsing preferences (Frazier, 1979; Frazier & Fodor, 1978; Kimball, 1973) that ignore semantic content on the first pass. Frazier (1979) formulates several of these, including the following two which are widely accepted in one form or another:

(6) *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)

(7) *Late closure* When possible, attach incoming material into the clause currently being parse (Frazier, 1979, p. 20)

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

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

(8) 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 (8 a) crashes, as it isn’t possible to incorporate the final word fell in a grammatical way. Reanalysis is required, with the grammatical parse being (8 b) where the matrix subject is *the horse raced past the barn*, a noun phrase (NP) containing a reduced relative clause *raced past the barn*. Thus *fell* can be incorporated as the matrix verb, with a structure comparable to, “The horse (that was) raced past the barn was hungry.”

There is an ongoing debate in the literature about what parsing model best fits the empirical facts. This study follows (Frazier & Clifton, 1996) in assuming that structure-first parsing strategies are at play, in addition to a primary vs. non-primary relation distinction that determines how immediately a phrase must be incorporated into a parse, allowing for some material to be incorporated later and thereby make use of additional information that is not available for immediate parsing decisions.

2.1 Structural overview of the ambiguity relevant to this study

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

(9) Jed crammed the newspapers under the sofa in the trashcan.

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

b) ✓ ... [$_{\text{VP}}$ crammed [$_{\text{NP}}$ the newspapers [$_{\text{PP1}}$ under [$_{\text{NP}}$ the sofa]]
[$_{\text{PP2}}$ in the trashcan]]
“#” indicates a structure with an implausible reading

In parsing (9 a), there is a fairly strong bias (due to *Minimal Attachment*, or some variation thereof), which favors a structure where the first PP attaches into the verb phrase (VP) as an argument of the verb, i.e. [$_{VP}$ V NP PP1], which leaves nowhere for the second PP to attach but as a modifier of the noun phrase (NP) inside PP1 ([$_{PP1}$ under [$_{NP}$ the sofa [$_{PP2}$ in the trashcan]]]). This initial parse (9 a) is pragmatically implausible, as one does not generally find sofas inside wastebaskets. Reanalysis is required to bring about the correct parse (9 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 it is *the newspapers under the sofa* that are being *crammed in the trashcan*.

Note that *Minimal Attachment* as defined by (Frazier, 1979) is somewhat at odds with recent developments in syntactic theory, e.g. obligatory binary branching (cf. Chomsky, 2014, p. 62). As originally postulated, *Minimal Attachment* relies on a verb with multiple internal arguments incorporating each of those arguments as a sister (i.e. a ternary branching structure: [$_{VP}$ V NP PP]). With current theories where binary branching is obligatory, two XPs (NP and PP) cannot both be syntactic sisters of the verb, so it becomes less clear that the VP attachment site for PP1 actually creates fewer nodes than the lower NP attachment site. Nonetheless, the preference for VP attachment in these kinds of sentences is there, be it due to Minimal Attachment, a preference for arguments over non-arguments, or something else, as evidenced by experimental data from e.g. Rayner et al. (1983) and Clifton et al. (1991).

This study is focused on the impact of speech act (interrogative vs. declarative) and its interaction with a trailing sequence of prepositional phrases (PPs), where the second is of two possible types. The contrasting types of PP2 shown in (10) are (a) a PP2 which must be an argument, and (b) one which can be a modifier.

(10) **PP2 types**

(a) *PP2 Argument* (Arg)

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

(b) *PP2 Modifier* (Mod)

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

Note that PP1, *in the drawer*, is the same in both (10 a) and (10 b), and is ambiguous on first encounter, as it could modify the paperwork or it could be the goal of cram. In (10 a), however, PP2 must ultimately be interpreted as the goal of the verb cram because into his briefcase cannot modify the drawer. Because cram only accepts one goal, this means that PP1 in (10 a) has to end up as a modifier of the paperwork. In (10 b), on the other hand, PP2 of his filing cabinet can (in this case, must) modify the drawer, and so in the drawer can and does end up as the goal of cram. The difference in PP2 status between (10 a) and (10 b) results in different structures, which I argue are reached by different parsing mechanisms. Namely, (10 a) should, by hypothesis, result in a parse which initially incorporates PP1 as the goal argument of cram but then fails and triggers reanalysis when PP2 is encountered. Conversely, (10 b) should by hypothesis allow a straightforward parse where PP1 is initially and ultimately slotted in as the goal of cram, since PP2 poses no issue when interpreted as a modifier of the drawer. This means (10 b) should not trigger reanalysis. Where (10 a) is a so-called garden path sentence, (10 b) is not.

In order for the differing parsing process for (10 a) and (10 b) to be explained by a strictly structurally based model of parsing, certain assumptions would have to be made about the syntax. A simple way to get the explanation to work is to assume that all arguments of a verb are syntactic sisters to the verb, resulting in a three-way branching VP for ditransitive verbs. In this case, in order to avoid postulating extra nodes that would be required for PP1 to be a modifier, *Minimal Attachment* dictates that PP1 should be assumed to fill the argument slot. This is not how modern syntactic theory assumes the structure looks, as three-way branching is proscribed. This paper is not interested in the particularities of syntactic theory, and it also is not necessary to rely on *Minimal Attachment* to make the necessary distinction, though it very likely does play a role. Instead, we can focus on distinction added to parsing theory in Construal (Frazier & Clifton, 1996): that of primary vs. non-primary relations.

The impetus behind adding this additional machinery to the theory of parsing is independently motivated: while structure-first decision making seems to hold for the parsing of many structures, there are some that seem to flout them. Construal illustrates this by way of relative clause (RC) attachment in constructions like (11).

(11) [$_{NP1}$ The daughteri] of [$_{NP2}$ the colonel$_j$] [$_{RC}$ who$_{i/j}$ was standing on the

balcony] …

The RC in (11) can modify either NP1, the daughter, or NP2 the colonel. A structure-first parsing system, together with the widely agreed upon structural parsing strategy Late Closure, would be expected to manifest as a consistent preference for the local attachment of the RC in (11), i.e. the structure where the RC modifies NP2. Instead, what Frazier and Clifton describe, based on a number of studies (e.g. Clifton, 1988; Cuetos & Mitchell, 1988) is a pattern where the preferred structure depends on the relationship between NP1 and NP2. They describe five categories of relationship, and a gradient of preferred RC attachment, from NP1 preference to NP2 preference.

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

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 [$_{\text{RC}}$ that was falling apart]

e) *Nonaccompaniment* with

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

Frazier & Clifton (1996) report that (12 a-b) type configurations favor NP1 RC attachment, (12 e) type configurations favor NP2 RC attachment, while (12 c-d) are intermediate. They argue that this gradient cannot be readily explained by structural parsing, and instead make use of a mechanism they call structural association. RCs are, 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 (12), the last theta assigner is NP2, and its domain extends up to the DP that contains NP1. This is a looser parsing decision that allows the syntactic structure to be decided on later, after semantic information becomes available: the RC can ultimately modify whichever member of the thematic domain is appropriate.

The crucial issue that distinguishes cases where structural association vs. structural parsing is appropriate is the idea of primary vs. non-primary relations. Frazier and Clifton formalize this distinction as (13).

(13) Primary phrases and relations include (Frazier & Clifton, 1996, p. 41)

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

b) Complements and obligatory constituents of primary phrases

RC attachment undergoes association because the relationship between a modifier and whatever is modified is a non-primary relation, and a relative clause is by definition a modifier and not an argument. Circling back to the PP-attachment that this study is concerned with, the argument vs. modifier distinction is precisely what distinguishes the two possible statuses of PP2 shown in (10) and repeated here as (14).

(14) **PP2 types**

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

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

Without locking down the exact syntactic structures that (14) represents, we can nonetheless say that the parser would seek to immediately incorporate PP1 into the tree in both cases. The infinitival (-finite) clause headed by *cram* is a primary phrase, and so its obligatory constituents hold primary relationships with *cram*. *Cram* takes an obligatory goal argument, so the parser cannot wait for semantic information to inform its association, it must make its best guess based on the principles of structural parsing, and attach it as an argument, as that property is what is forcing the immediate decision to be made. When PP2 is encountered, reanalysis will be required, in the case of (14 a), or it will not be, in the case of (14 b).

2.2 Interrogativity

The focus of the current study is to examine the impact that interrogativity has on the reanalysis just described, motivated by the observation that for many speakers the interrogative version of a sentence like (16) is easier to process than its declarative counterpart (15).

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

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

The question that must be asked, then, is what exactly differs between (15) and (16)? Syntactically, very little: the position of the subject he and the auxiliary had have been reversed.

Semantically, or perhaps it is better to say pragmatically, there are a number of differences, which I will perfunctorily discuss. The details may not be quite write, as the focus of this paper lies elsewhere, but it's important to be aware of the general ideas presented.

The pragmatic differences between (15) and (16) lie with the presuppositions the sentences carry with them, and with the placement of focus. The declarative in (15) has few presuppositions beyond the existence of the actors and objects involved (the referrants of *he*, *paperwork*, *drawer* and *briefcase*), and that these actors and objects can be involved in *cramming*. The presuppositions of (16) are a super set of those of (15): a yes/no question additionally presupposes that the listener knows the answer to the question, for one. Further presuppositions might exist, depending on where the focus lies within the sentence.

Focus in a declarative like (15) is typically broad, meaning no element is having attention called to it. A polar question like (16), however, will typically receive narrow focus on one element, so that when uttered, one element is more prominent than the others. The focused element becomes the part of the sentence that the question is about. Focus can fall on any of the lexical or referential elements (subject, matrix verb, infinitival verb, object NP, or the NP of either PP1 or PP2) of the sentence, or the auxiliary verb.

(17) **Had** he planned to cram the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his briefcase]?

(18) Had **he** planned to cram the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his briefcase]?

(19) Had he **planned** to cram the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his briefcase]?

(20) Had he planned to **cram** the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his briefcase]?

(21) Had he planned to cram the **paperwork** [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his briefcase]?

(22) Had he planned to cram the paperwork [$_{\mathrm{PP}_1}$ in the **drawer**] [$_{\mathrm{PP}_2}$ into his briefcase]?

(23) Had he planned to cram the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into **his** briefcase]?

(24) Had he planned to cram the paperwork [$_{\mathrm{PP}_1}$ in the drawer] [$_{\mathrm{PP}_2}$ into his **briefcase**]?

In (17), with focus on the auxiliary, the question is about the entire proposition, and whether or not it is true. In this case, there are not any additional presuppositions. In (18), with focus on *he*, the question is asking about whether the referent of *he* is the actor who performed the action described; in this case, the entire predicate is presupposed: someone *had planned to cram the paper in the drawer into his briefcase*, but was it *him*? Skipping ahead to (22), with focus on *drawer*, the question is instead about which *paperwork* this is all happening to: *the paperwork in the drawer*, or some other stack of paperwork? In this case, it is presupposed that the referent of *he* was the one who *had planned to cram some paperwork into his briefcase*, and only the exact referent of *the paperwork* is not presupposed. For each other location of focus, the presuppositional content is similarly complementary to whichever element is focused and therefore being asked about.

This set of pragmatic differences between (15) and (16) might very well be the source of the intuition that (16) is easier to comprehend than (15), but that is not the only possibility. Another significant difference between the two speech acts is the prosody and intonational melody. While dialects of English differ, there is typically a difference in melody between a declarative and question, and in many American English dialects, the interrogative is pronounced with a final rise, while the declarative exhibits just a series of downsteps. This difference is the one that the current study explores, to see if it can readily explain the intuitive difference in processing difficulty.

2.3 Prosody of questions vs. declaratives

In pursuing the possibility that it is the intonation and prosody of polar interrogatives which creates the intuitive contrast that this study investigates, we must consider what question intonation actually sounds like. It is generally agreed that in American English, the intonation of a polar (yes/no) question has the property of a final rise. Indeed, this has been confirmed in corpus studies such as Hedberg, Sosa, & Görgülü (2017) who found that 79.8% of the 410 American English yes/no questions in their study (ten-minute phone conversations from the CallHome Corpus of American English and the Fisher English Corpus) had a "low-rise nuclear contour" (L*H-H%, L*H-☐H%, or L*L-H%)[1]. To briefly explain their ToBI notation, a tone T is either L for low or H for high; T* is anchored to the stressed syllable, and T- and T% are boundary tones (intermediate phrase boundary and intonational phrase boundary respectively). See, e.g., *Guidleines for ToBI labeling* (Beckman & Ayers, 1997) for a more thorough explanation of ToBI. An additional 10.7% of the Hedberg et al. (2017) data had a "high-rise nuclear contour" (the authors categorizes the following tunes as "high-rise nuclear contours:" H*H-H%, or !H*L-L%). That leaves only 9.5% spread across the other 5 categories (High-fall, Rise-fall, Low-fall, Fall-rise, and Level). Only 5.6% of the data showed a falling contour. According to the authors' analyses, these contours occur on the final main stress of a sentence and thereafter. In the case of the types of sentences examined in the current study,that would result in a rising contour on

[1] Hedberg et al. (2017) use ☐ to indicate an upstep, which is not standardly transcribed with ToBI.

the head noun of the final PP.

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

The need to prepare for that 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 high attachment.

A brief informal survey found that most speakers maintain low tones on prior stresses, although some had a H tone on the subject noun. It also varied between speakers and between sentences as to whether there is a prosodic boundary (marked by a low tone and/or pause) immediately before the rise (after PP1) or not.

The prosodic structures found in the data collected for the current study are discussed in 4.2.

2.4 Can prosody affect parsing?

A number of studies have shown that in listening to speech, prosodic cues appear to help reduce the frequency with which incorrect parsing (i.e. a garden path) occurs. For example, Kjelgaard & Speer (1999) conducted a study using digital manipulation of recorded speech to create three versions of sentences containing a garden path temporary ambiguity (discussed above). They recorded speakers saying sentences with natural prosody, such as the following pair:

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

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

They then cross-spliced these together to make several versions. One version had prosodic cues which cooperated with the intended reading of the sentence; another attempted to have "neutral" prosody; and the third used intentionally misleading prosody. The initial fragment of each was then presented to participants (the portion from the beginning of the sentence to the word house in (26 - 27) and they were asked to agree or disagree with whether a visually presented word, either *is* or *it's* was likely to be the next word in the sentence. Participants gave more accurate and speedier judgements when the prosodic cues lined up with the correct parsing. The results of this study, as well as a growing body of literature, suggest that that prosodic information can (or perhaps must) be used by the parser in making processing decisions.

Consider Fodor (2002) analysis of relative clause attachment preference. This concerns sentences such as (28):

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

The relative clause (RC) who was on the balcony can attach either locally (low, modifying N2), making it *the actress* who was *on the balcony*, or higher up (non-locally, modifying N1), so that we understand *the servant* to be the one who was *on the balcony.* In these sorts of sentences, Cuetos & Mitchell (1988) found a 60% preference for low attachment in English speakers, but only a 40% preference for low attachment in Spanish speakers. In apparent violation of the general preference for local attachment, some languages, like French and Spanish (and

Russian, but not Romanian or Brazilian Portuguese, so this is not a general feature of Romance languages), prefer to attach relative clauses higher, while others more often obey Late Closure (e.g., Swedish, Egyptian Arabic, and English). This non-local preference is weakened in cases where the ambiguous RC is short (one prosodic word). Fodor (2002) asserts that these tendencies exist in both listening to spoken words (under conditions where a particular parse is not favored by the explicit prosody) and in silent reading.

Fodor notes that other researchers have shown the presence and absence of prosodic breaks to influence parsing decisions, and specifically that the presence of a prosodic break before the RC in sentences like (28) encourages high attachment. Fodor leverages this in order to explain the difference in RC attachment site tendency between languages. She argues that the phenomenon can be neatly account for by linking attachment site preference to the likelihood of a prosodic break before the RC. This difference in prosodic tendency, in turn, can be explained using a constraints-based approach. Consider Selkirk's (1986) alignment constraints:

(29) Align(αCat, E; βCat, E)

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

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

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

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

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

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

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

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

2.5 Predictions for the current study

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

(32) SUBJ V OBJ PP1 PP2

I suggest that high attachment of PP2 will be marked by a prosodic boundary between PP1 and PP2 (for discussion of what constitutes a prosodic boundary, see e.g. Streeter (1978) and Salverda, Dahan, & McQueen (2003)). Low attachment of

PP2, on the other hand, will lack any substantial boundary marking.

(33) *Hypothesis 1*

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

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

(34) *Hypothesis 2*

A first reading (no preview) of a GP sentence will exhibit less natural prosody (more hesitation at and after the disambiguating region) than:

- A first reading of a non-GP sentence.
- A second reading of a GP sentence.

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

(35) *Hypothesis 3*

A first reading of a garden-path sentence will more often be produced with prosodic structure that represents an implausible or ungrammatical parse of the string (low attachment of PP2), whereas a second reading sentence will more often be pronounced with the prosodic structure that represents the intended parse (high attachment of PP2).

Note that hypothesis 3 can't be applied in cases where the reader fails to successfully and fluently produce the sentence (which may sometimes happen, due to the garden-path effect in the Arg condition).

Finally, I investigate an intuition originally discovered by Dr. Janet Fodor and Dr. Dianne Bradley: that these GP sentences are not as difficult to parse when encountered in interrogative, as opposed to declarative, context.

(36) *Hypothesis 4*

Reading 1 of a declarative GP sentence 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 Readubg 1 of an interrogative Arg sentence.

I revisit these hypotheses in the results chapter and the discussion and conclusion chapter.

Chapter 3

Methodology

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

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

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

3.1 Participants recruitment

All participants in the were undergraduate students enrolled at Queens College in Psychology 101[1] who participated for course credit. Self-reported age ranged from 18 to 25 years. Participants were recruited a software system designed for university participant pools. Students saw a recruitment notice on the system website (see Appendix A), 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.

[1] IRB approval number: 2018-0072

3.2 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 4.3.1 of the results chapter).

3.3 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[2] was done using the Open Sesame software (Mathôt, Schreij, & Theeuwes, 2012) which provides a graphical user interface, scripting language, and interpretation of Python code. The system was capable of 10-20 millisecond accuracy, with the display's 60Hz refresh rate being the limiting factor. Key input had a latency of about 10ms.

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

[2] Instructions were also provided verbally and via printout, see appendix B.

3.4 Materials

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

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

Table 3.1: Illustrative experimental item, constructed in four versions

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

The experimental stimuli were based on earlier pilot study exploring this same phenomenon Peckenpaugh (2016), with several adjustments made to accommodate the objectives of the current study. The sequence of parts for each of the basic items was always the same, shown in (37).

(37)

	complex verb cluster					
introductory material			*the construction*			
Subject	Auxiliary	Matrix verb	Infinitival verb	Object	PP1	PP2
He or She	*had*	*mental state verb:*	*cram, put, stick, or set*		*always ambiguous*	*potential disambiguation*
*Order shown for **D** condition; reversed in **Q** condition*					*in/on/ under*	***Arg:*** *into/onto or* ***Mod:*** *of/from*

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*, *want*, or *plan*).

The verb within the construction did not vary within a quadruple, but a given quadruple could have one of four verbs: *cram*, *put*, *stick* or *set*. 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 stick*, *decided to put*, and *decided to set*); 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 3.1's illustrative example). That is, PP1 was temporarily ambiguous in every version of a given quadruple, being interpretable as either the goal argument of the construction verb or as a modifier of the object NP. However, in Arg versions of a quadruple, the argument interpretation of PP1 cannot be sustained once PP2 is encountered. In those cases PP2 must fill the goal argument slot and PP1 must be a modifier. The working assumptions about parsing discussed earlier, i.e., that the parser will initially assume PP1 to be the goal argument due to the primary status of arguments, assumes that Arg versions of a quadruple require reanalysis. Between quadruples, the preposition that headed PP1 varied, but was always one which was compatible with it being a goal argument or a modifier of the object: *in* (8), *on* (7), and in one case, *under*.

One benefit of using a complex verb cluster (auxiliary + matrix participle + infinitive) rather than a single verb[3] was that it isolated the differences across the versions of a quadruple triggered by the speech act factor to the left extremity of the introductory material of the sentence: only the position of the subject and the

[3] 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 ...*?)

auxiliary were affected, meaning that the construction itself was completely untouched by this manipulation.

The purpose of including introductory matrix verbs was to reduce the oddity of the polar interrogative versions of each quadruple. It seems odd to ask, “Did Mary put the jelly beans in the window onto a fancy dish?” because, when it is clear that the speaker already knows so much about the situation, it becomes difficult to imagine a pragmatically plausible context where such a question would be asked. Such sentences might well be described as “prosecutorial[4].” Arguably, this is somewhat mitigated by the addition of a verb like *decided*: rather than asking about facts that we already seem to know, we are instead asking about an actor’s mental state with regard to those facts. Even if we know the facts of the situation, we do not necessarily know, for instance, whether it was the result of a decision, some third party’s action, or mere happenstance. Another adjustment made in order to make the polar interrogative versions of each quadruple more pragmatically acceptable limited the amount of detail in the experimental sentences, so that fewer adjectives and adverbs were included compared to the items employed in Peckenpaugh (2016), and subjects were always third person nominative pronouns (*he* or *she*).

Importantly, the construction verb was always one which demanded a goal argument. Where some of the verbs used in the items employed by Peckenpaugh (2016) only optionally took a goal, the current study used only verbs which require a goal argument. Verbs that optionally take a goal might result in a parse where

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

PP1 is not immediately incorporated as the goal argument, which would mean that PP2 would not necessarily force reanalysis. Consider the contrasting sub-categorization of the verbs in (38) and (39):

(38) **Optional goal** (*hide*)

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

✓ The gangsters had hidden the shotguns.

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

Another important consideration was ensuring that the Arg versions had a PP2 which definitively disambiguated the attachment site of PP1 such that reanalysis was forced. In (40), PP2 is implausible as a modifier of *rocking horse*, but not strictly impossible, and the sentence is grammatical with PP2 modifying it. On the other hand, the use of *onto* in (41) completely disallows the modifier interpretation of PP2 at the syntactic level: a PP headed by *onto* cannot grammatically modify the preceding NP.

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

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

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

It is worth noting that some linguists (e.g. Den Dikken (2006)) believe *of* is not a preposition in the same sense as *from*, *on*, or *in*, etc., in that it appears to be serving a strictly grammatical or functional purpose, without real lexical content. Importantly, it is also only 2 characters, whereas *into*, *onto*, and *from* (the other possible heads of PP2) are all 4 characters. This is revisited and its possible impact is explored in the results section (section 4.3.3).

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 (*decided*, *intended*, *wanted*, or *planned*) to support more pragmatically plausible questions. PP1 was always interpretable as either the goal argument or a modifier of the object. PP2 differed across the PP2-Status factor, but not across the speech act factor. In the two Arg versions of a quadruple, it was headed by *into* or *onto* and was intended to force reanalysis, under the assumption that PP1 had been incorporated into the parse as an argument, since a PP headed by *into* or *onto* must be interpreted as the goal argument, the position that PP1 would have presumably been occupying in the ongoing parse. For the two Mod versions of a quadruple, PP2 was headed by *from* or *of* and therefore was not expected to require reanalysis, as *from*- and *of*-headed PPs can attach as modifiers of a preceding NP (in this case, the NP within PP1), allowing PP1 to stay in the goal argument slot.

3.4.2 Fillers

There were 32 filler items that ranged in complexity: [[TODO some were ordinary sentences, some contained center embedding, and others included attachment ambiguities]]. 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). A full list of fillers is available in Appendix D.

All filler items had the same sort of introductory material as the experimental

Table 3.2: Illustrative experimental item, constructed in four 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?

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

3.4.3 Length

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

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

3.5 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).

3.6 Procedure

Participants were given a verbal overview of the experimental procedure and then asked to read a one page printout of the procedure before signing a consent form. After signing a consent form, participants sat at the computer and were once again walked through instructions before the first practice item was presented.

Participants sat at a computer and used keyboard button presses to navigate the experimental presentation. They received thorough instructions and completed three 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, in order to allow some time for the participant to settle into the procedure before any results were recorded.

Each experimental item was preceded by a screen showing a line of ten Xes with its left edge aligned with the left edge of the to-be-revealed sentence. This was designed the participant's attention on the start of the sentence, and hopefully avoid unintended look-ahead. The issue of potential look-ahead is discussed at greater length in section 3.6.1.

The fixation screen remained visible indefinitely, until the participant pressed *START*.

After *START* was pressed, recording of the first reading began and the sentence appeared on the screen in black font on a light blue background. The recording continued and the screened remained visible until the participant pressed *NEXT*.

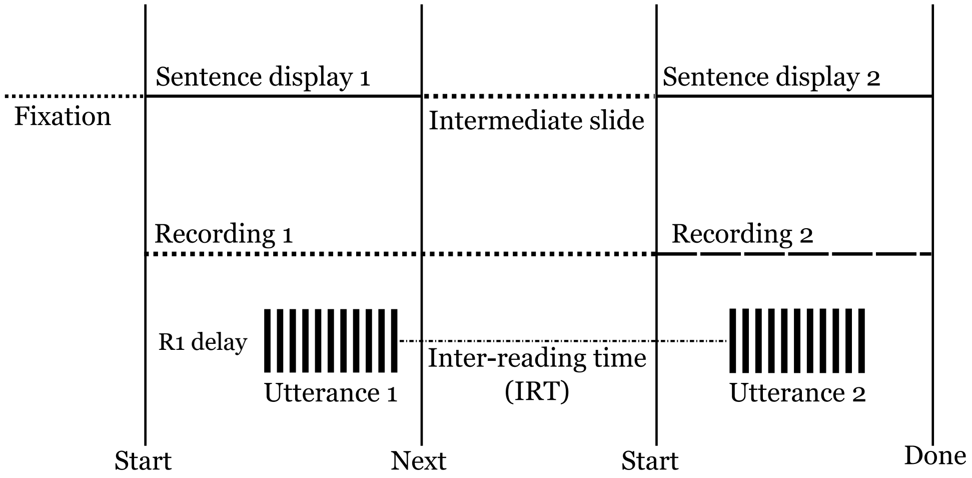


Figure 3.1: Procedure diagram

After pressing *NEXT*, a screen appeared with instructions telling the participant that they were between readings, and needed to press *START* to reveal the sentence again and prepare for their second reading. Immediately after *START* was pressed, the first recording ended and the second recording began, and the sentence reappeared, this time on a light green background. There were never any linebreaks in item display.

The shifting of required key presses and the changing background color were intended to aid the participant in remember where they were in the process, and to prevent accidental double-presses of any given button from having unintended side effects. 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.

3.6.1 On look-ahead

An advantage that the Double Reading Procedure has is that it allows for certain assumptions to be made about Reading 2 that otherwise would be unclear: Reading 2 certainly represents a *considered* reading of the sentence. Not only has the reader had ample time to examine the sentence, but has necessarily read it and heard it read in producing Reading 1. This means Reading 2 can plausibly be thought to represent a considered prosodic structure, at least more so than an entirely naive reading, and should not reflect any processing issues; a parse should have already been devloped 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 delievered without preview. The properties of these Reading 1 delays are discussed at length in a later section, but for now it suffices to say that the very limited preview is possible during a delay that typically falls in the 0.2 to 2.7s range (median = 1s, SD = 0.4). As an example of common reading rates, Ashby, Yang, Evans, & Rayner (2012) reported faster readers as averaging 328 words per minute (wpm), and slower readers 228wpm, in silent reading. That study found that reading time is slower for reading aloud, and that the availability of parafoveal information (i.e., the difference between 1 word and 3 word windows) is less impactful for that reading mode. Given that the experimental items range from 13 to 15 words, most of the R1 delays would not allow even a fast reader to read the entire sentence: the median 1s R1 delay would allow a fast reader time to read very few words; keep in mind that the window is even shorter, because in addition to just reading, the subject is also handling several other cognitive processes (e.g., visual processing, lexical access, issuing motor commands, etc.). The utterance of Reading 1 should, therefore, contain within it any behavioral reflex of whatever parsing difficult the reader has, for most recordings.

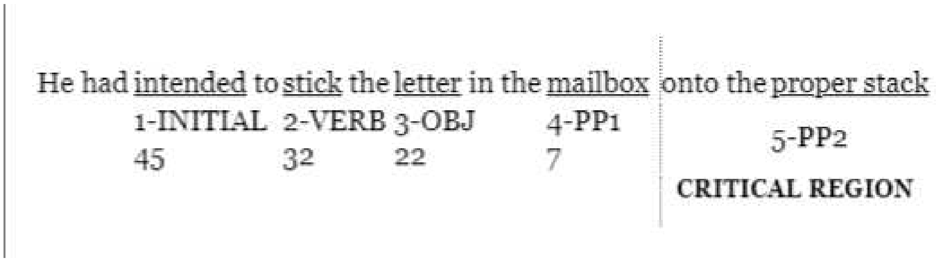
In order to clearly understand the results of this double reading study, it is important to understand the mechanics of reading. Specifically, we would want to know at what point during the reading of a temporarily ambiguous sentence the participant will become aware of the existence of a disambiguating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

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

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

The relevant question for the study at hand is as follows: how much of the sentence will the reader have seen and processed when a given word is being spoken? A typical item is displayed in (42), 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.

(42)



He had intended	to stick	the letter	in the mailbox	onto the proper stack
1-INITIAL	2-VERB	3-OBJ	4-PP1	5-PP2
45	32	22	7	CRITICAL REGION

Table 3.3 describes these distances across items all experimental items. Note that these values do not vary across condition, because counting starts after both the subject and auxiliary verb, and ends before PP2, and the only changes across versions are subject-auxiliary inversion and the content of PP2.

Table 3.3: Distance in characters from fixation to disambiguation in experimental items

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

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

When does the reader become aware of the existence of PP2? When fixated on the direct object head noun, the range of distance is 21 to 27 characters, with a median of 25.5: PP2's content is still outside of view, even in the case of the smallest distance, and adjusting it to be a few characters smaller to account for the fact that fixation is likely to occur closer to the center of a word rather than on its first character. At most, the presence of the first few characters of PP2's preposition may be available, but certainly not the character space after it. The distance from the PP1 fixation point (the head noun within that PP) ranges from 5 to 9 characters, with a median between 7 and 8 characters. Thus, we can say with some certainty that the reader of a sentence such as (42) will be aware that another phrase, one which starts with a 4-character word, remains to be incorporated into the parse sometime after processing of the direct object, and before processing of PP1.

There is yet another piece to consider: the so-called eye-voice span (EVS), and the fact that the readers in this study are reading aloud rather than silently. According to Laubrock & Kliegl (2015), when reading aloud the voice is typically behind the eyes by some 10-20 characters (M = 16.2 characters, SD = 5.2 characters). Adjusting Table 3.3 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 3.4.

Table 3.4: EVS-adjusted character distance to disambiguation in experimental items

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

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

3.7 Inter-reading time (IRT) measurement

Subjects were asked to read each sentence twice, once with no preview at all (Reading 1), and then again after unlimited preview (Reading 2). Inter-reading time (IRT) is a measure of the amount of time between when a participant stops speaking after a cold reading and when speaking resumes for a previewed reading.

IRT was measured using a Python script and Google's WebRTC Voice Activity Detection (VAD) over 44.1kHz WAV files down-sampled to 8kHz via SOX[5]. This VAD system uses Gaussian Mixture Models to make probabilistic decisions as to whether a given audio frame is speech or noise (see Falk & Chan (2006) for a complete description). Google's implementation takes one parameter, which they call aggressiveness: a 4-tier setting for the level of confidence necessary to call a given frame speech. I call this "rejection rate", where a higher rejection rate means that the model requires a high level of confidence before assuming a frame is speech, i.e., it is more likely to label something noise than speech. The implementation codes this setting as 0-3, 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 measurement of the least modified data that gave plausible

[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.

measurements. Specifically, each file was measured using the highest possible rejection rate for the VAD algorithm and no modification of the file. 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 rejection rate for the VAD 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.

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

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, as Goldman-Eisler (1961) found that a large majority (82.5 to 87%) of pauses in fluent speech are less than 1s. Stimuli range from 18-22 syllables in length. If we assume a speech rate of 3 to 7 syllables per second (Jacewicz, Fox, & Wei, 2010) we would expect utterances between 2.5s and 7.3s. Conservative thresholds higher and lower than the expected were used, especially on the higher end, to allow for any difficulties processing or fluency that might have lead to longer reading times.

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.

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

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 4.3.1).

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

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

3.8 Prosodic judgments

A trained linguist informant naive to the research being conducted listened to recordings and reported the presence or absence of breaks in certain regions of the sentence, as well as several other judgments. She was instructed to familiarize

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

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 Table 3.5, as well as full plain-text lists of all items.

Table 3.5: Sentence region labels

SUBJ		V		OBJ	PP1	PP2
He	had	meant	to stick $\|\|_{V}$	the pencil case $\|\|_{OBJ}$	in the cabinet $\|\|_{PP1}$	into his book bag.
NP$_{SUBJ}$	AUX	V$_1$	V$_2$	NP$_{OBJ}$	PP$_1$	PP$_2$

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

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

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

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

3.8.1 Reliability

A second trained linguist repeated the task over 120 recordings selected from 8 participants (two from each group, one per ordering). Even number experimental items were used from 4 participants, and odd numbered from the other 4. There were 8 recordings missing from the 128 selected, so the reliability task resulted in judgments over 120 recordings. The first informant also blindly re-rated those 120, with the recording name obscured and instructions not to revisit her original ratings. Reliability scores (percent of recordings agreed upon) are reported in Table 3.6.

Table 3.6: Inter and intra-rater agreement

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

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

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 judgements over three break locations (the third, V, is omitted throughout this report because it was extremely rare, occuring in just over 8% of recordings). As such, disagreement on that break and the fact that break strength is actually a compilation of two judgements (weakest and strongest break) amplified the noise to some extent.

Chapter 4

Results and discussion

This section reports various descriptions and analyses of the recordings obtained. The reported results include the effect of Speech Act (declarative/D vs. interrogative/Q) and PP2 Status (argument/Arg vs. modifier/Mod) on the location of prosodic breaks, as well as on time spent reflecting upon a sentence between readings, which I call inter-reading time (IRT). In order to evaluate the extent to which participants adhered to the protocol as intended, i.e., began to read immediately for Reading 1 as opposed to producing a considered reading in Reading 2, the delay for which a sentence is displayed before a participant begins to read it is compared for Reading 1 (R1 delay) vs. Reading 2 (R2 delay[1]). The prosodic patterns for participants with especially fast and especially slow R1 delays

[1] Note that R2 delay is subsumed by but not synonymous with IRT; IRT includes lag time after R1 is completed but before the recording for R2 begins. This distinction is laid out in more detail in Section 4.3.2.

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.

4.1 Data for analysis

Data for 32 total participants were analyzed. Given 4 versions of the experiment and 2 possible orderings there would ideally be 4 participants per version-order combination. Ultimately, 3 participants had to be excluded for different reasons, resulting in the distribution is as shown in Table 4.1[2]. 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 4.1: No. of participants per version-order combination

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

Some of the expected 3072 recordings (32 participants x 48 items (16 experimental and 32 filler) x 2 readings) were not used due to intrusive noise

[2] 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.

duringthe recording session. Additionally, data were also excluded from analysis if any (Reading 1/Reading 2 pair) was missing; there were 9 such incomplete pairs excluded. Without analyzable data from both members of a pair, it is difficult to determine the extent to which the elicitation protocol was executed as intended (i.e., the extent of preview for Reading 1 vs. Reading 2).

For experimental items, 978 recordings were subjected to prosodic analysis, constituting 95.6% of the utterances elicited. Because IRT data considered utterances in pairs (Reading 1/Reading 2) rather than separately, the database for response timing took in 489 datapoints.

Table 4.2: Number of recordings analyzed, as a function of Speech Act and PP2 Status

	D	Q
Arg	244	240
Mod	246	248

4.2 Prosodic break patterns

In what follows, the distrbution of OBJ breaks and PP1 breaks are reported as a function of the four sentence types created by the materials design (D/Q x Arg/Mod), for each of Reading 1 and Reading 2. Then, the patterns of breaks over the two positions are considered, before moving to statistical analysis. Note that while breaks after the construction verb were reported, these breaks were exceptionally rare and occured in only 8% of recordings, so they have been set aside. The break locations are indicated with a % symbol in (43).

(43)

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

As noted in section 3.8, the results reported are based on the subjective judgements of a trained linguist who was naive to the purposes and hypotheses underlying the research.

4.2.1 Individual break patterns

Table 4.3: Percent occurance of OBJ break (frequency of occurence in parenthesis) as a function of sentence type and Reading

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

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

Table 4.4: PP1 break by condition and reading

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

The PP1 break was almost always present for cases where PP2 was an argument;

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

4.2.2 Combined break patterns

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

Table 4.5: Percent occurence of both breaks as a function of sentence type and Reading

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

Table 4.5 shows that for Arg sentences, there are very few instances with the OBJ-only pattern (0.8% in declaratives, 2.5% in interrogatives); whereas that pattern is fairly frequent for Mod sentences (31.1% in declaratives, 31.4% in interrogatives). The pattern with both breaks is somewhat more common for Arg sentences (72.1% in declaratives, 71.7% in interrogatives) than Mod (54.1% in declaratives, 43.0% in interrogatives). The PP1-only pattern occurs at about the same rate in Arg interrogatives (25.8%) as in Mod interrogatives (25.6%) and Arg

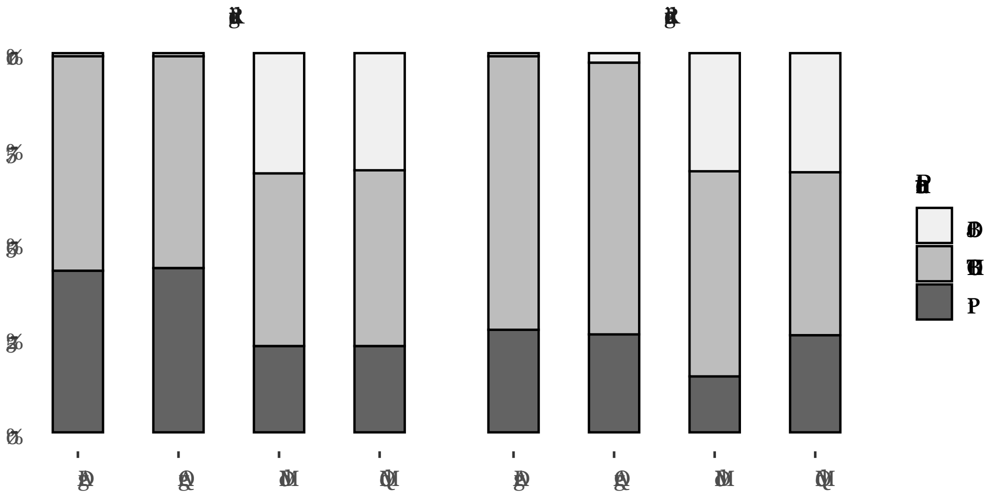


Figure 4.1: Break pattern as a function of sentence type and Reading

declaratives (27%), but is noticeably less common for Mod declaratives (14.8%). These proportions are visually represented in figure 4.1.

4.2.3 Break Dominance

The relative strength of the PP1 and OBJ breaks was also collected. Figure 4.2 incorporates this information, where “PP1 dominance” means that the PP1 break was reported to be stronger than the OBJ break; “OBJ” dominance means the opposite; and “Equal strength” means that neither break was reported to be stronger than the other (the 5 instances with no breaks were again omitted).

When looking at the combined break patterns, one can think of there being three bins: the PP1 bin, the OBJ bin, and a neutral bin between them. In Section 4.2.2,

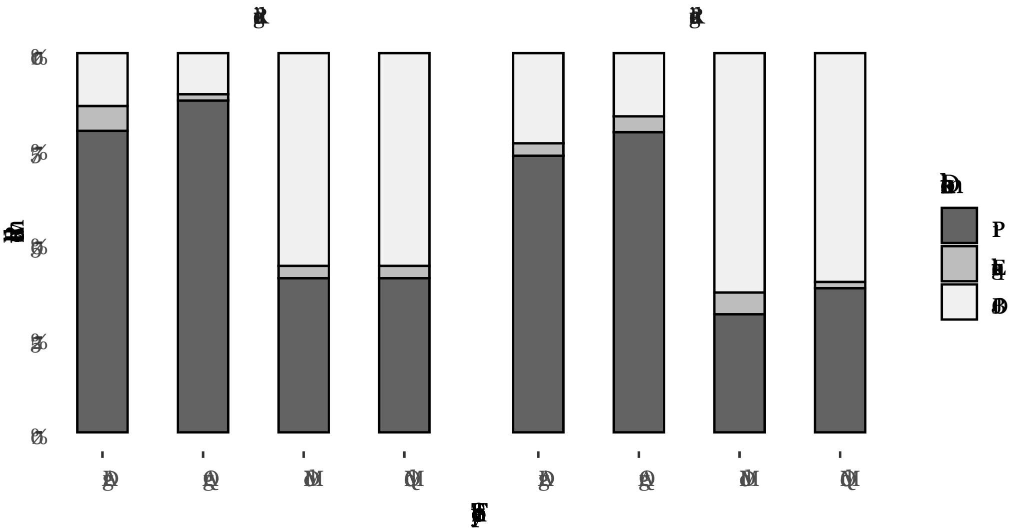


Figure 4.2: Percent break dominance occurence as a function of sentence type and Reading

the neutral bin containing instances of both breaks occuring is robust. This break dominance analysis distributes most of those cases that have both breaks into either the OBJ or PP1 bin, depending on which break is more prominent. When the breaks are of equal strength, they remain in the middle bin, but there are much fewer such cases when looking at dominance instead of simple occurence.

Figure 4.2 clearly shows a robust effect of PP2 status on break dominance, and little to no impact of reading or speech act.

4.2.4 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 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). When 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.

In each case, the model with the lowest reported AIC was selected from a set of models. Model comparisons did not always find significant differences between the more complex models. In each case, the selected model was compared to a minimal model where where fixed effect variables were removed, leaving only an intercept, and all reported models represent improvement over the minimal model to a statistically significant degree. That comparison is reported for each model. All regression models were run using the lme4 R package (Bates, Maechler, Bolker, & Walker (2019)), with p-values calculated via the lmerTest R package (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen (2019)).

4.2.4.1 Break occurence

Table 4.6 shows a model predicting the occurrence of an OBJ break with estimates for the coefficients of the fixed effects of Reading 2, PP2 and the interaction between Reading and PP2 status. 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, χ^2(2)=30.5, p < 0.001).

Table 4.6: Mixed effects logistic regression model predicting OBJ break occurrence

Outcome: OBJ break	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 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, χ^2(1)=228.0, p < 0.001).

[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(¬A)). A log odds of 1.39 translates to an odds ratio of 2.4:1 (1.39^e=2.4) and a probability of 71% (2.4/(1+2.4)=0.71).

Table 4.7: Mixed effects logistic regression model predicting PP1 break occurrence

Outcome: PP1 break	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.

4.2.4.2 Break dominance

Models were also run for predicting break dominance. Table 4.8 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 4.8: Mixed effects logistic regression model predicting OBJ break dominance

Outcome: OBJ dominance	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

Table 4.9 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.

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

Table 4.9: Mixed effects logistic regression model predicting PP1 break dominance

Outcome: PP1 dominance	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

sentence with argument PP2s than sentences with modifier PP2s, with interrogatives having slightly increased log odds of PP1 break dominance. Log odds of PP1 break dominance were slightly less in Reading 2 than Reading 1.

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 4.10: Mixed effects logistic regression models predicting break dominance in Reading 2

	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 at the p < 0.05 level, but its inclusion marginally improves the fit of each model.

4.2.5 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. As a way of analyzing the protocol, and the extent to which participants performed as expected, participants were categorized based on their median R1 delay. In what follows, a fast median R1 delay was shorter than or equal to 0.9s, and a slow one was longer than 1.05s, resulting in 12 participants per category. Ten participants had R1 delays between those values, categorized as "normal," and ignored. These calculations were done over Reading 1 of experimental items (n = 489).

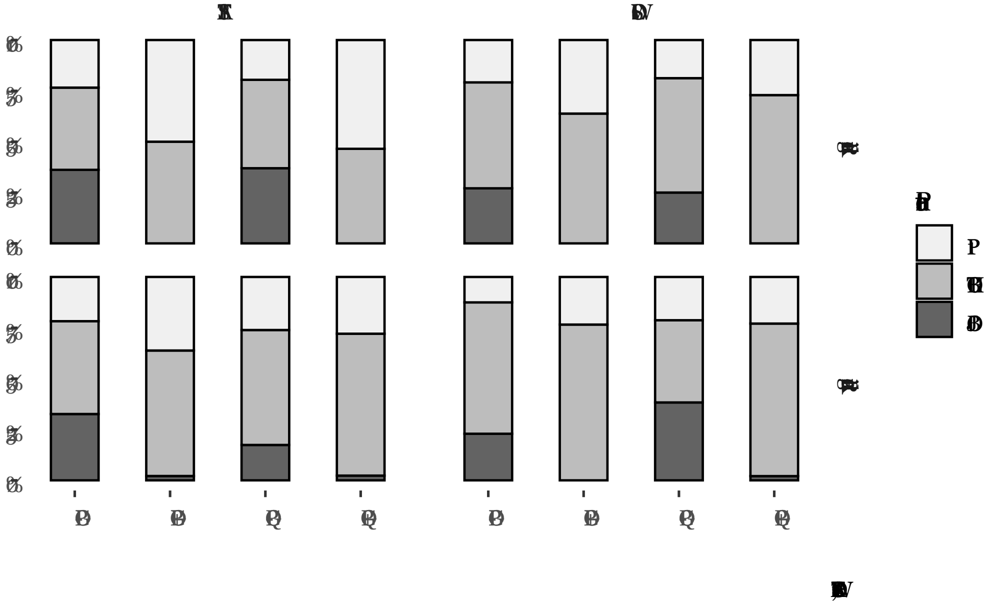


Table 4.11: Simple break pattern by condition and R1 delay category

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

There appears to be a clear difference between the participants categorized as having fast or slow R1 delays in the extent to which they used both breaks, with the slow category of participants preferring more breaks than the fast category of participants. There also seems to be a larger increase in usage of both breaks from Reading 1 to Reading 2 for the fast category than for the slow.

TODO add hypothesis testing and more detail for
the above statements

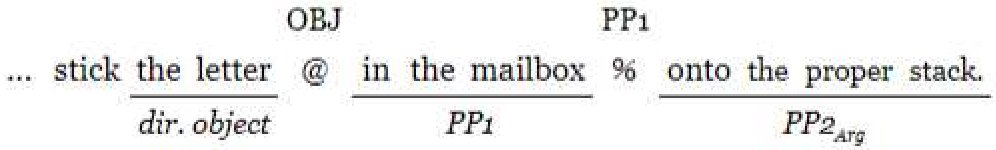

OBJ
PP1
... stick the letter @ in the mailbox % onto the proper stack.
dir. object
PP1
PP2Arg

Figure 4.3: ... stick the letter @ in the mailbox % onto the proper stack

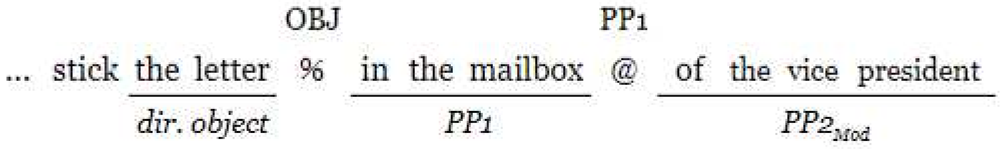

OBJ
PP1
... stick the letter % in the mailbox @ of the vice president
dir. object
PP1
PP2Mod

Figure 4.4: ... stick the letter % in the mailbox @ of the vice president

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

Essentially, the expected patterns can be described as in (44) and (45), where % represents a robust prosodic break, and @ represents a weaker break or no break.

(44)

(45)

This is supportive of hypothesis 1 presented in the introduction; and, in fact, of a broader formulation of it.

Hypothesis 1

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

Broadened hypothesis 1

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

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

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

Hypothesis 2

A first reading (no preview) of a GP sentence will exhibit less natural prosody (more hesitation at and after the disambiguating region) than: * A first reading of a non-GP sentence. * A second reading of a GP sentence.

Hypothesis 3

A first reading of a garden-path sentence will more often be produced with prosodic structure that represents an implausible or ungrammatical parse of the string (low attachment of PP2), whereas a second reading sentence will more often be pronounced with the prosodic structure that represents the intended parse (high attachment of PP2).

It is difficult to know whether a given reading represents more or less natural prosody, but given that there's a difference, it seems most likely that Reading 2 is the more natural of the two.

It is surprising that the effect of PP2 status is generally lessened in Reading 2 when compared to Reading 1, but this can likely be explained as an epiphenomenon. 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's 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's perceived as a dominant break after PP1 in the PP2 argument condition, when they're confused, that effect of PP2 status will go away in Reading 2 once they've had time to figure the sentence out. This might mean that the noise caused by readers that are simply breaking phrase-by-phrase is actually amplified in Reading 2.

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

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

4.3 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 3.7.

TODO Explain what I see as the importance of
IRT

4.3.1 Data cleanup

IRTs below 0.25s (2) and above 25.0s (5) were assumed to be implausible and omitted. Experimental data were then Winsorized by participant to bring data below the 2.5% and above the 97.5% threshold to the value at those thresholds. The resulting measure is referred to as wIRT and is distributed as shown in figure 4.5 (n = 489).

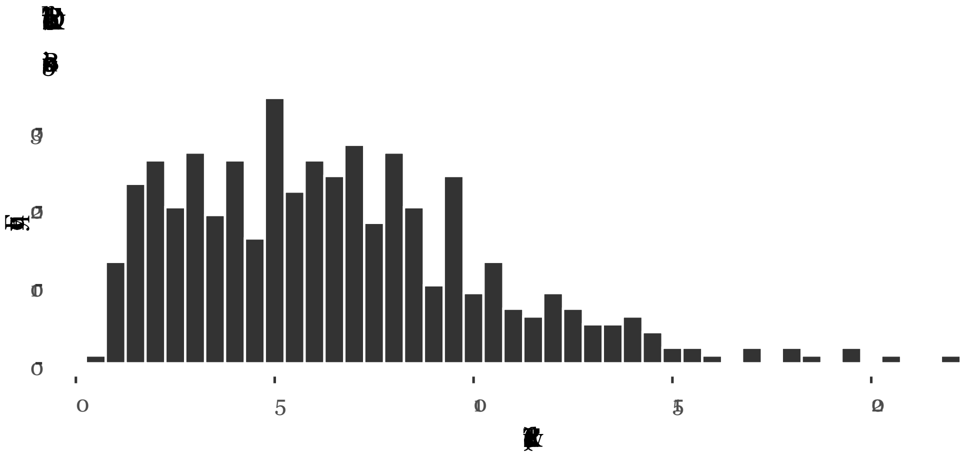


Figure 4.5: Distribution of wIRT

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.

4.3.2 IRT results

Table 4.12 shows the mean wIRT for each experimental condition. While mean wIRT increases from modifier PP2 to argument PP2 more for declaratives than for interrogatives, the difference in that increase is very small (0.04s).

Table 4.12: Means (s) by condition

Condition	D	Q	Q - D
Mod	6.20	6.47	0.27
Arg	6.61	6.85	0.24
Mod - Arg	0.41	0.37	-0.04

The same numbers are represented visually in figure 4.6

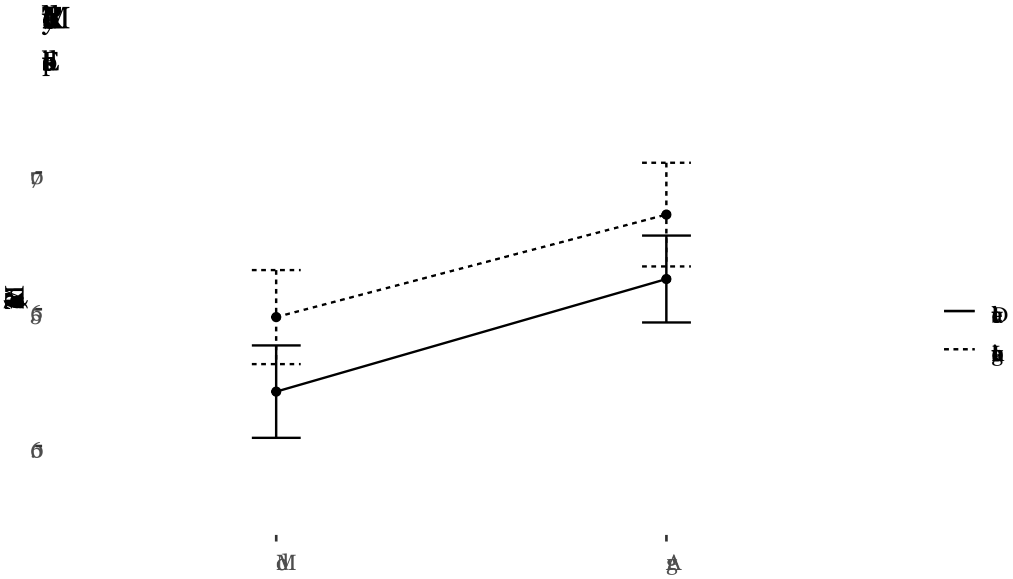


Figure 4.6: Mean IRT by condition

The two slopes are only very slightly divergent. Notably, both speech act and PP2 status appear to have main effects on wIRT, with interrogatives having longer IRTs than declaratives, and sentences with argument PP2s having substantially longer IRTs than those with modifier PP2s.

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

By hypothesis, the best model for predicting IRT should be one with fixed effects of speech act and PP2 stats and the interaction between them. This model is shown in table 4.13.

Table 4.13: Linear mixed effects regression model predicting wIRT by condition with interaction term

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

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

Table 4.14: Linear mixed effects regression model predicting wIRT by condition

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

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

4.3.3 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 condition, while half used from. In the Arg condition, half used into and half used onto to head PP2. An analysis that looks at the identity of the PP2 head found that while into and onto do not behave differently, of and from do. Starting from a very 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, except it substitutes the lexical identity of the PP2 head for the Arg vs. Mod categorization, with into and onto collapsed into one level of the PP2 factor and used as the reference level.

Table 4.15: Linear mixed effects regression model predicting wIRT by speech act and PP2 head

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

Sentences where PP2 was headed by from typically had wIRTs that were 0.84s faster than sentences where PP2 was headed by into/onto; when the PP2 head was from wIRT was only 0.08s faster than for into/onto.

4.4 Discussion of IRT results

It's clear from the above that argument PP2s (ones headed by into/onto) result in longer wIRT measures. It also appears that interrogativity increases wIRT to a lesser extent, regardless of the PP2 status. Because the interaction between the two factors is not a significant predictor of wIRT, we are left to assume that either wIRT does not represent a behavioral reflex of the intuition that interrogativity makes difficult to process PP2-attachment ambiguities easier, or else the current study does not have enough power to detect it.

The difference between from and of PP2s is also potentially a source of noise. The from sentences are less clearly disambiguated than the of sentences. Where (3) has only one reading, (4) has another possible reading, albeit somewhat implausible: i.e., we could imagine that in (3) from her brother-in-law modifies the cookies, while in (4) of the minivan cannot modify the bicycle.

She had intended to put the bicycle on the roof rack of the minivan. She had decided to cram the cookies in the basket from her brother-in-law.

This lingering ambiguity could have increased wIRT somewhat because the reader, given unlimited time, spent some of that time noticing and then eliminating that possible reading. The difference here can be explained by once again making an appeal to structural parsing vs. structural association: if we imagine that a from PP is associated rather than parsed, the reader is free to consider other possible interpretations. Because of can be seen as less a preposition and more a functional head, it stands to reason that it would be treated different, as something that must be parsed immediately, i.e. that it instantiates a primary relation.

It could also be a simpler explanation: the fact that of is only two characters, whereas from and into/onto are all four characters, participants may have recognized a pattern, wherein they could be sure that a two character PP2 head meant the sentence did have the difficult properties of some of the ones with four character PP2 heads (most notably those pesky into/onto ones), and eliminated some of the needed study time.

A study where the modifier PP2s are more tightly controlled with regard to the head preposition, both in terms of the length and the properties of the head, might produce clearer results.

4.5 The processing cost of interrogativity

It is worth taking note of the fact that the mean wIRT for interrogative versions (6.7s) of the sentences in the reported study was longer than for the declaratives (6.4s). Similarly, Peckenpaugh (2016) found that reading times for interrogatives were longer than for declaratives. This is perhaps representative of the processing cost of interrogativity reported by Mehler (1963) who found that a so-called kernel sentence, i.e. a simple declarative, was easier to recall verbatim than was a number of sentences that he considered to be syntactic transformations of that kernel sentence (K): negative (N), polar question (Q), passive (P), and combinations thereof: NQ, NP, QP and NPQ. Mehler found that accurate recall was more frequent for K sentences (300/460, 65.2%) than for the other sentences types, with interrogatives (210/460, 45.7%) being recalled accurately at a lower rate than the two other individual transformations (234/460, 50.9% for N; 243/460, 52.8% for P).

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

Interrogative status itself appears to increase the time needed for participants to feel they've 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 5

Discussion and conclusion

This section will go through the motivating questions behind this study and discuss the extent to which those questions are answered, or not. It will discuss the issues that may have lead to the disatisfying. It will then go on to present further questions and propose further studies to explore those new questions and the ones left unanswered here.

5.1 The intuition

Recall that the primary motivator for this study was the observation that PP-attachment garden paths appear easier to understand for many speakers of American English when presented in the interrogative, as opposed to the declarative. One goal of this study was to find a behavioral correlate of this intuition. Another was to seek an explanation for the intuition itself.

5.1.1 Behavioral correlate for the intuition

Ultimately the data here are unable to provide a clear answer to whether or not IRT represents a behavioral correlate of the intuition that PP-attachment garden paths are rendered easier to parse when presented as polar questions. While the difference in mean IRT for Q+GP compared to Q-GP was 0.04s less than for D+GP compared to D-GP, showing a smaller increase in processing time for interrogatives than for declaratives, this is a rather small difference when the grand mean is 6.6s. As such, mixed-effect regression analyses were not able to detect stastical significance for the interaction between interogativity and ±GP. That said, it was also not clear that a model without the interaction was a better fit than one with it (Full model AIC = 9103.5; Non-interaction model AIC = 9101.6; $X^2 = 0.035$; $p > 0.8$). This does not, of course, negate the intuition; it simply means that we have not yet found a behavior that can be said with certainty to correspond to that intuition.

There are a number of issues that might be preventing the observation of behavior that confirms the intuition, and there are also possible explanations for why the intuition might exist without a behavior correlate. The possibilty most generous to the reported study is that the sample size here is simply inadequate. It's possible that with more data, this very procedure would have provided a stastically significant behavioral correlate of the intuition.

It is also possible that the experimental paradigm used here is too lenient to capture the behavior it is looking for: it relies on participants to begin Reading 1 without any lookahead simply by their own recognizance rather than any experimenter-enforced mechanism. Perhaps some participants were able to, perhaps unintentionally, get enough information a priori to negate the effect of the interaction on their IRT.

It could also be that IRT is simply the wrong measure for detecting a behavioral correlate of the intuition. The reanalysis triggered by +GP Reading 1 could very well be all that was needed to make the understanding of +GP Reading 2 trivial; this seems unlikely, though, because a significant effect of +GP on IRT was found.

5.1.2 Explaining the intuition

A possible explanation for the intuitive effect of interogativity on parsing garden paths is provided in the work by Bader (1998). Bader demonstrates that it is easier to recover a 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 intuition this study is concerned with, this would mean that should sentences be more similar across the garden path vs. non-garden path condition in the interrogative than in the declarative, the intuitive reduction in difficulty of reanalysis would naturally follow.

While the data presented above do not definitively show that the prosodic structure of questions is less different for questions than it is for declaratives, it also does not rule it out. Because there does not appear to be a one-to-one mapping of prosodic structure to syntactic structure in the recordings, one cannot make a clear-cut declaration of what a given sentence type looks like prosodically. Instead, we must observe the percentage of recordings for each version of the sentences that exhibit a given pattern. The relevant table is repeated here as 5.1 for convenience.

Table 5.1: Combined breaks per condition in reading 2 only

	Non-garden path		Garden path	
	D	Q	D	Q
Both	54.1	43.0	72.1	71.7
OBJ	31.1	31.4	0.8	2.5
PP1	14.8	25.6	27.0	25.8

While there is a larger drop in the number of utterances with both breaks from +GP to -GP for questions than for decaratives, the opposite is true for the pattern with a PP1 break by itself. There is an argument to be made that the OBJ break, which does not correspond to a change in syntactic branching direction for any version of the sentences, is a hesitation rather than a break, due to a moment of confusion or perhaps simply a need to breathe.

To be more explicit, for this explanation to work, the prosodic structures in 46 and 47 would likely be the ones considered “correct” or most common for a the declarative versions. The symbol “%” is being used to represent a prominent prosodic break.

(46) *D+GP*: He had crammed [$_{\text{OBJ}}$ the newspapers] [$_{\text{PP1}}$ under the sofa] % [$_{\text{PP2}}$ into the trashcan].

(47) *D-GP*: He had crammed [$_{\text{OBJ}}$ the newspapers] under the sofa in the

guestroom.

The prosodic structure of 47 does not mandate any major break, because there is no change in syntactic branching directory and no other reason to mandate a break there. That is, the PP1 break differentiates the two syntactic structures and signals the attachment site of PP2.

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 a tonal change, which might sound very similar to a prosodic break. The symbol "//" indicates the start of the rising contour.

(48) *Q+GP*: Had he crammed [$_{\text{OBJ}}$ the newspapers] [$_{\text{PP1}}$ under the sofa] %+// [$_{\text{PP2}}$ into the trashcan]?

(49) *Q-GP*: Had he crammed [$_{\text{OBJ}}$ the newspapers] [$_{\text{PP1}}$ under the sofa] // [$_{\text{PP2}}$ in the guestroom]?

The critical issue here is that in (49), the absence of a syntactically-motivated break is hidden by the juncture created by the start of the final-rising contour of a question.

If the PP1 break is treated as the main indicator of the prosodic structure, then the larger difference across ±GP for declaratives (12.2%) than for interrogatives (0.2%) does indeed, though perhaps weakly, leave open the door for the Bader (1998) style explanation of the intuition.

5.2 On future work

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

5.2.1 Hesitations vs. prosodic breaks

A major roadblock to the success of the methodology here reported is that the region of the sentence where the critical prosodic difference between +GP and -GP sentences lies is the same region where disambiguation occurs for the +GP sentences. Some weaknesses in the data as they are leave need for future studies and methodological improvements.

A major issue is the difficulty in differentiating between syntactically-motivated prosodic breaks and hesitations. As table 5.1 and the reliability data reported earlier in this paper illustrates, a linguistically trained judge will not necessarily be able to distinguish between prosodic breaks and hesitations, and will not necessarily agree with another linguistically trained judge on the relative prominence of breaks and hesitations within a recording. It is famously difficult to instrumentally describe a prosodic break, and there is no guarentee that examining the wave forms would fare any better in making this distinction than the intuition of the judges. That said, it may be possible to distinguish the two sorts of breaks: it may be that boundary tones or segmental lengthening are typically not part of a hesitation break, but are a part of hesitation break. No immediately apparent pattern is available, but it would be inordinately beneficial if one could be found.

In 5.2.3 I propose that an event-related potential paradigm might hold the key for distinguishing between the two.

5.2.2 Embedded questions

The explanation for the intuition that PP-attachment garden paths are less difficult to parse in interrogatives than in declaratives must either be prosodic, or else semantic/pragmatic. It seems impossible that the very minor syntactic difference across the two sentence types (subject-auxiliary inversion) could be the reason for a difference in percieved ease of understanding. It is either the difference in the intonational contour between the two sentences types, or else it is something about the meaning or metalinguistic difference between interrogatives and declaratives. The study just reported has not provided adequately convincing evidence that the inuition can or can not be explained entirely by the prosodic differences between questions and declaratives. It seems that the 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. For example:

(50) *EmQ +GP:* He asked her if she had decided to cram the old newspapers under the couch in the wastebasket.

(51) *EmQ -GP:* He asked her if she had decided to cram the old newspapers under the couch in the guestroom.

(52) *EmD +GP:* She told him that she had decided to cram the old newspapers

under the couch in the wastebasket.

(53) *EmD -GP:* 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 (50), does not differ from that of a sentence like (52); but, the semantic properties and some of the pragmatic properties of the embedded clause in (50) are the same as in one of the Q+GP sentences from the study reported here.

(54) Had she decided to crammed the old newspapers under the couch in the wastebasket?

These stimuli could be used in another reading aloud study, like the one just reported, or could be used with ERP or eye-tracking methodology. If a behavioral correlate of the intuition is found with embedded question stimuli, that would be strong evidence that the explanation is semantic or pragmatic rather than prosodic.

5.2.3 Event-related potential (ERP)

An event-related potential study of the phenomenon could provide a number of useful insights. It has been shown that the so-called 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. This has a number of important implications for research on the phenomena at hand.

For one, it provides a way to distinguish between a hesitation after PP1 and a true prosodic break; which is to say, a way to distingiush between an unencumbered reading with a prosodic break after PP1 and a garden-path effect that, in the study reported here, resulted in a hesitation at that point. An ERP measure time-locked to that position should show a CPS component when the reader succesfully incoprorates PP1 as a modifier of the object and then changes branching direction to incorporate PP2 as the goal argument; conversely, a reader who is experiencing a garden-path effect at that point should instead or additionally show a P600 component (Steinhauer (2003) found that the two components can be additive). If the participant were concurrently recorded reading the sentence aloud, definitive data could be created which could then be used to either train a linguist or perhaps a neural network to distinguish between hesitations (P600s) and prosodic breaks (CPSes). The trained model or linguist could then return to the recordings collect for this study and make definitive judgements regarding the nature of the breaks reported here, and clarify the results of this study.

It would also be useful to see if repeated readings of the same garden-path sentence reduces the P600; if so, is the rate of decay faster for interrogatives than declaratives? Such a phenomenon could be a correlate of the reported intuition[1].

[1] QUESTION: Do P600 components reflect the strength of a garden path? Also, do I need to go into any detail here about how ERP works or what the properties of P600 and CPS are?

Appendix A

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 B

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 B.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 C

Experimental items

Version	Text
D Arg	She had decided to cram the cookies in the basket into her jacket pocket.
D Mod	She had decided to cram the cookies in the basket from her brother-in-law.
Q Arg	Had she 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 Arg	She had decided to put the child on the rocking horse onto the see-saw.
D Mod	She had decided to put the child on the rocking horse from his parents.
Q Arg	Had she 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 Arg	He had decided to set the board games on the floor onto the card table.
D Mod	He had decided to set the board games on the floor of the living room.
Q Arg	Had he 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 Arg He had 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 Arg Had he 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 Arg He had intended to cram the paperwork in the drawer into his boss's desk.

D Mod He had intended to cram the paperwork in the drawer of his filing cabinet.

Q Arg Had he 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 Arg He had intended to put the bicycle on the roof rack into the garage.

D Mod He had intended to put the bicycle on the roof rack of the minivan.

Q Arg Had he 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 Arg She had intended to set the clothes in the hamper onto the dresser.

D Mod She had intended to set the clothes in the hamper from his sister.

Q Arg Had she 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 Arg She had intended to stick the letter in the mailbox onto the proper stack.

D Mod She had intended to stick the letter in the mailbox of the vice president.

Q Arg Had she 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 Arg She had planned to cram the stolen files in the wall-safe into a suitcase.

D Mod She had planned to cram the stolen files in the wall-safe of their hideout.

Q Arg Had she 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 Arg She had planned to put the jelly beans in the window onto a fancy dish.

D Mod She had planned to put the jelly beans in the window of his candy store.

Q Arg Had she 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 Arg He had planned to set the appetizers on the platter onto the buffet.

D Mod He had planned to set the appetizers on the platter from his cousin.

Q Arg Had he 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 Arg He had planned to stick the post-it note on the handout onto his notebook.

D Mod He had planned to stick the post-it note on the handout from the lecture.

Q Arg Had he 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 Arg He had wanted to cram the newspapers under the sofa into the wastebasket.

D Mod He had wanted to cram the newspapers under the sofa from the thrift store.

Q Arg Had he 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 Arg He had wanted to put the photo on the coffee table onto the mantelpiece.

D Mod He had wanted to put the photo on the coffee table from his grandfather.

Q Arg Had he 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 Arg She had wanted to set the textbooks on the top shelf into the file box.

D Mod	She had wanted to set the textbooks on the top shelf of the book shelf.
Q Arg	Had she 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 Arg	She had wanted to stick the golf clubs in the back room into the closet.
D Mod	She had wanted to stick the golf clubs in the back room of their condo.
Q Arg	Had she 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?

Appendix D

Filler items

Version	Text
-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 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 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	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	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 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 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 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 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 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 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 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 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 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 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 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 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 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 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?

+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 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 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 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 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 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 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 circulate the latest job posting his company had sent.

-PP Q Had he remembered to circulate the latest job posting his company had sent?

+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 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 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 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 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?

Appendix E

Instructions to RA

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:

Table E.1: Sentence region labels

SUBJ		V		OBJ	PP1	PP2
He	had	meant	to stick $\|\|_V$	the pencil case $\|\|_{OBJ}$	in the cabinet $\|\|_{PP1}$	into his book bag.
NP$_{SUBJ}$	AUX	V$_1$	V$_2$	NP$_{OBJ}$	PP$_1$	PP$_2$

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.

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.

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

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.

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

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

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.

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.

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.