Working memory resource allocation is not modulated by clausal discourse status

Margaret Kroll*, Matt Wagers
Department of Linguistics, University of California, Santa Cruz

ABSTRACT
It is well-established that increasing the length of a dependency within a sentence, such as by adding material between a subject and verb, increases the processing difficulty of the sentence. Recent research by Dillon et al. (2014, 2017) shows that length effects are not equally observed between appositive clauses and restrictive relative clauses, with restrictive relative clauses showing greater sensitivity to additional material than appositive clauses. The authors suggest that the divergent behavior is due to the differing at-issue statuses of the two clause types. We present acceptability judgment data inconsistent with the hypothesis that the length effect differences are due to the at-issueness of the two clause types. We also give acceptability judgment data testing the hypothesis that implicit prosody plays a role in the length effect observations. We conclude by suggesting that additional research be undertaken to explore the role of attention modulation between the two clauses.

Keywords: parentheticals, relative clauses, appositives, working memory, implicit prosody, discourse status

Introduction

A main puzzle in psycholinguistics is determining what makes some clauses more difficult to process than others. One property that has been tied to processing difficulty is length: longer clauses require more processing power and lead to greater processing difficulty than shorter clauses. However, there is debate over how to best characterize this generalization and over what the right function is that relates length to processing difficulty. Research has shown that there is no single mapping from clausal length to level of processing complexity. For example, recursive right-branching structures can embed multiple clauses without greatly increasing processing difficulty, while center-embedded structures are challenging to process beyond one or two recursions (Eady & Fodor, 1981; Frazier, 1985). Additional research has shown that longer clauses, defined in terms of string length, increase processing difficulty when intervening in a dependency, such as between a wh-phrase and its gap (Gibson 1998; Hale 2001; Warren & Gibson 2002; McElree et al. 2003; Van Dyke & Lewis 2003; Grodner & Gibson 2005; Lewis & Vasishth 2005; Levy 2008; Bartek et al. 2011). These findings suggest that length increases processing difficulty because it furthers the distance between two elements that form a dependency.

This paper investigates an additional case in which clausal processing difficulty does not straightforwardly map onto clausal length. As mentioned, researchers have shown that increasing string length between two elements that form a dependency increases processing difficulty. This increase in processing difficulty is referred to as a length effect. Length effects are attributed to the preference in natural language to minimize the linear length of dependencies within a sentence, e.g. by minimizing the number of words between a wh-word and its gap, or between a subject and verb (Frazier, 1985; Hawkins, 1994, 2004; Gibson, 1998, 2000; King & Just, 1991; Gibson & Pernlmutter, 1994; Grodner & Gibson, 2005; Temperley, 2007; Gildea & Temperley, 2010). In several recent experiments, Dillon et al. (2014, 2017) document unexpected experimental results in which the discourse status of an embedded clause appears to contributes to clausal complexity. Specifically, they find length effects even when the clause does not intervene within a dependency, and furthermore suggest that processing difficulty is driven not only by clausal length—re-
gardless of intervention within a dependency—but also by clausal discourse status.

Consider examples (1) and (2). From the point of view of dependency-intervention the examples are the same: the distance from the subject to the verb in each sentence is the same. However, from the point of the discourse, the examples are distinct. Indeed, Dillon et al. found that lengthening a restrictive relative clause, such as by adding in the bolded material in (1), leads to a greater decrease in the acceptability of the containing sentence than lengthening an appositive (parenthetical) clause in a corresponding sentence, such as in (2).

(1) The man who was on the cruise Mary took to the Pacific Islands tried to throw a waitress over board.

(2) The man, who was on the cruise Mary took to the Pacific Islands, tried to throw a waitress over board.

Dillon et al. suggest that the length effect difference between restrictive and non-restrictive relative clauses provides evidence that the processing difficulty of the clause types is driven in part by the discourse status of the clauses. The hypothesis that discourse status affects the allocation of processing resources is supported by findings of research on discourse entities, in which the discourse status of noun referents has been shown to impact the processing of NPs. Researchers have given experimental evidence that, for example, introducing a new discourse referent imposes a processing cost (Murphy, 1984; Haviland & Clark, 1974), that the choice of referring expression is partially determined by the cognitive status of the referent in the mind of addressees and speakers (Gundel et al., 1993; Ariel, 1990), and that the givenness status of NP referents that intervene within a dependency affects the difficulty of processing that dependency (Gibson, 1998, 2000; Warren & Gibson, 2007). Dillon et al. propose that restrictive relative clauses, as contributing at-issue content to their containing sentence, contribute a greater processing cost than parenthetical clauses, which contribute not-at-issue content to their containing sentence.

An explanation for how discourse status can affect processing difficult can be found in research on the properties and constraints of the memory system used in language processing. Insights into the nature of this memory system are provided by domain-general cognitive psychology models of cognitive architecture (Anderson, 2007), as well as by linguists’ implementation of these models for language comprehension (Lewis & Vasishth, 2005; i.a.). These architectures provide frameworks for understanding why some sentences, such as those containing wh-dependencies or center-embeddings, are more difficult to process than others, with explanations provided in terms of interference and decay. Although the specifics of different models diverge, there is wide agreement that sentence processing is constrained not only by the grammar, but also by domain-general cognitive mechanisms, such as a capacity-limited short-term memory for linguistic meaning and form. Dillon et al. propose that memory architecture is differentiated based on the at-issueness of linguistic material. Specifically, they propose that not-at-issue discourse content draws from separate working memory resources than at-issue discourse content. Therefore, appositives, as not-at-issue content, draw from separate working memory resources than their surrounding context. Restrictive relative clauses, on the other hand, express at-issue discourse content and, as such, draw from the same memory resources as matrix clause content.

We present experimental evidence that challenges a view in which the length effect difference between restrictive and non-restrictive relative clauses is due to the at-issueness of the clause types. Drawing on recent corpus and experimental work by Harris & Potts (2009), Simons et al. (2010), AnderBois et al. (2011), Tonhauser (2012), Koev (2013), and Syrett & Koev (2015), we show that manipulating the discourse context so that appositives contribute at-issue discourse content does not modulate the length effect difference between appositives and restrictive relative clauses. We conclude it is therefore unlikely that discourse status is the factor behind the length effect difference. We additionally investigate the alternative hypothesis that the observed length effect differences are driven at least partially by the implicit prosody projected by readers onto experimental items during silent reading. Specifically, we propose that the length effect differences are in part attributable to implicit prosodic factors that affect how structures are chunked in short term memory during reading. This account is
compatible with one interpretation of the claims of Dillon et al., namely that the two clause types differentially consume working memory resources. However, under our proposal, working memory resources are consumed according to prosodic phrasing, instead of differentiated by discourse status.

**At-issue and not-at-issue content**

Since Potts (2005), appositives have been understood to contribute a special kind of not-at-issue meaning called *conventional implicature* meaning. As conventional implicatures, appositives have distinct properties from main clause at-issue content. Potts locates the nexus of difference at the level of semantic composition: appositives compose semantically independently from main clause content, and therefore contribute a distinct proposition at the semantic level. Because appositives compose independently from main clause content, they are inaccessible for binding from a main clause and generally project past negation and propositional attitude verbs (Karttunen, 1974; Karttunen & Peters, 1979; Beaver, 2001; Kadmon, 2001; Amaral et al., 2007; Potts, 2005/2012; Simons et al., 2010).

Subsequent research also suggests that appositives differ from at-issue content at the discourse level: appositives generally contribute secondary information, and they are less likely to be targeted by direct rejection and to serve as the antecedent for ellipsis (Syrett & Koev, 2015). One main diagnostic for characterizing not-at-issue content is that it generally does not contribute toward resolving the current Question Under Discussion (QUD) of the discourse (Ginzburg, 1996; Simons et al., 2010; Roberts, 1996/2012). For example, the appositive in (3) expresses information that is secondary to the QU D, which is resolved, or answered, in the main clause.

(3) **Q:** What are the sailors doing?
**A:** The sailors, who see the coming storm, are sailing back into the harbor.

An open question is how or whether the differences between appositive and at-issue content are represented in memory. Under the assumption that restrictive relatives and appositives differ in contributing at-issue and not-at-issue content, respectively, the recent findings by Dillon et al. (2014, 2017) that acceptability ratings are much more sensitive to the length of restrictive relatives than to the length of appositives support a model of discourse processing in which the parsing operations that construct at-issue and not-at-issue content proceed independently. Dillon et al. note that their results are compatible with both an interference-prone memory retrieval process (Van Dyke & Lewis, 2003; Bartek et al., 2011), under which at-issue and not-at-issue content are processed using different cues that distinguish the two types of content at retrieval, and with a theory in which at-issue and not-at-issue content are represented in different short-term memory buffers. If in different buffers the two types of content would not interfere with each other in short-term memory, as in for example competition over limited working memory resources. Because the length effects are found in both medial and final appositives, the latter of which does not intervene between a subject and verb dependency, the authors argue that a separate memory buffer explanation is more convincing than an interference or retrieval-based explanation.

For the current study, we draw from recent experimental and corpus evidence that has challenged the assumption that appositives always contribute not-at-issue content. AnderBois et al. (2011) and Syrett & Koev (2015) argue that appositives can sometimes be interpreted as containing at-issue content; we propose that this recent research suggests a way to tease apart clausal effects from at-issueness: if memory is sensitive to at-issueness, then the appositive and restrictive relative length effects should be mediated by factors that control the at-issueness of the appositives. That is, if the effects found by Dillon et al. are due to the not-at-issue discourse status of appositives, then an appositive that is interpreted as at-issue should show the same behavior as an at-issue restrictive relative clause. Note that while we follow the recent literature in taking the ability of appositives to sometimes contribute at-issue content, this is not a universally
held position. Potts (2012), for example, argues that appositives always contribute not-at-issue content as a result of their semantic and syntactic properties. While Potts agrees that appositives can sometimes behave as at-issue content, e.g. in their ability to answer a QUD, for him this ability does not change their inherent not-at-issue status. Under this definition, not-at-issueness is attached to particular constructions that differ at the level of semantic composition, not at the level of discourse. See Appendix for additional discussion.

**Implicit Prosody**

We propose an alternative explanation of Dillon et al.’s findings and the findings of the current paper. While appositives and restrictive relative clauses differ along the at-issueness dimension, the clauses also differ along the prosodic dimension. Namely, appositives are marked by a prosodic break at the beginning and at the end of the clause. When reading, these prosodic breaks are generated by the **implicit prosody** that is assigned to the punctuation delineating the appositive clause. Implicit prosody refers to the prosody that readers assign to text when reading silently. Theories of how implicit prosody affects processing are found mainly in research studying syntactic attachment ambiguities and wrap-up effects. Early observations that prosodic factors affect language processing during silent reading are due to Fodor, who proposed that sentence comprehension involves both a syntactic **and** a prosodic parser. The prosodic parser packages material into phrases roughly balanced in length, and the syntactic parser then interprets only those sentence parts that are packaged in a prosodic phrase. Fodor named this theory the Implicit Prosody Hypothesis (Fodor, 1998, 2002; Quinn et al., 2000):

**Implicit Prosody Hypothesis (IPH):** In silent reading, a default prosodic contour is projected onto the stimulus, and it may influence syntactic ambiguity resolution. Other things being equal, the parser favors the syntactic analysis associated with the most natural (default) prosodic contour for the construction.

Fodor proposes that the Implicit Prosody Hypothesis can account for why some languages, like English, have a low attachment preference (akin to Frazier’s (1978) Late Closure principle), while other languages, like Spanish, have a high attachment preference (Cuetos & Mitchell, 1988) in syntactically ambiguous sentences. For example, while English speakers tend to attach the ambiguous relative clause in (4) low, to the NP *the actress*, Spanish speakers tend to attach the clause high, to the NP *the servant*.

(4) Someone shot the servant of the actress who is on the balcony.

Fodor proposes that because the grammar of English does not dictate whether there should be a prosodic break before a relative clause, the length of the relative clause will affect whether or not the clause is prosodically incorporated into its preceding DP. Specifically, a shorter relative clause is more likely to not contain a prosodic break between its head DP than a long relative clause. When a prosodic break *does* occur, people are more likely to choose a low attachment reading (Fodor, ibid.; Jun, 2003; Swets et al., 2007; a.o.). Research has also shown that there is individual variation in the preference of how to prosodically parse a DP head with a modifying relative clause (Jun, 2003; Jun & Bishop, 2015), which has been attributed to individual differences in working memory capacity (Swets et al., 2007).

We hypothesize that there are two possible ways in which implicit prosody could be driving the length effects. The first is that the long restrictive relative clauses are more difficult to prosodically parse than the long appositives because the restrictive relatives lack prosodic boundary cues. This difficulty can lead to a higher processing cost of the sentences, and may also lead to infelicitous parses. For example, readers may initially attempt to parse the entire restrictive relative clause into one prosodic phrase; because the clauses are too large to be parsed into a single phrase, readers may be forced to break at an unnatural point within the clause (e.g. at a point at which there are no significant corresponding syntactic boundaries). Under this hypothesis, it is the prosodic ill-formedness of the sentence that directly leads to the decline in acceptability ratings.

1 Unlike, for example, the grammar of French, which Fodor claims to require a prosodic break in this location. French is also claimed to have a high attachment preference (Frenck-Mestre & Pynte, 2000).
Such a theory is supported by experimental evidence that punctuation can facilitate reading. For example, Staub (2007) gives eye tracking evidence that disruption occurs when a comma is absent at a subordinate/main clause boundary where readers expect a prosodic break, even if there is no syntactic ambiguity. For example, more reading disruption occurs at [the vet] in (5) than in (6), in terms of slower reading times and more regressions.

(5) When the dog arrived at the clinic the vet and his assistant went home.

(6) When the dog arrived at the clinic, the vet and his assistant went home.

Slow-downs also occur at [the clinic] in (6), which has been attributed to theories of wrap-up effects. For example, Hill & Murray (2000) argue that punctuation may give rise to a low-level hesitation response of the oculomotor system, because punctuation often signals increased difficulty. Hirotani et al. (2006) argue that clause endings with punctuation are marked by a specific contour of implicit prosody, which leads readers to pause at the end of the phrase.

Additionally, Warren et al. (2009) give eye tracking evidence that wrap-up effects are driven by punctuation, even when occurring sentence-medially. Their eye tracking measures showed a disruption at [phoned] in (7) compared to (8), with longer reading times and more regressions.

(7) Joe and Bob phoned, before leaving.

(8) Joe and Bob phoned before leaving.

Their hypothesis is that the slow-down is a low-level oculomotor response to punctuation, which allows time for additional processing (Just & Carpenter, 1980) but is not necessarily driven by the need for additional processing. Kuperman et al. (2010) has more recently argued that there are start-up effects and wrap-up effects over lines of text, the latter of which is probably driven by comprehension and semantic integration processes.

An alternative hypothesis is that one interpretation of Dillon et al.’s explanation is correct, namely that appositives differentially consume working memory resources from main clause content; however, the memory resource consumption is driven by prosodic phrasing and not by discourse status. Under this hypothesis, it is not the prosodic ill-formedness itself that is directly driving the rating decrease of the long restrictive relatives, it is the burdening of working memory that results from the ill-formedness that is directly driving the rating decrease. This hypothesis is supported by research such as that of Schafer (1997), which shows that in listening experiments, reanalysis of a word was more difficult (took longer) when the lexical item was in a separate intonational phrase than the disambiguating material. If the prosodic and syntactic parsers work in tandem, then we might think that prosodically grouped constituents will be more likely to share short term memory resources. A version of this system is argued in Hirotani (2004) and Hirotani et al. (2006). We found suggestive evidence that implicit prosody played a role in the length effects found in our experiments, and conclude that while implicit prosody is a contributing factor to the length effects, it is likely not the only factor.

The current study

Seven acceptability judgment studies provide findings that are inconsistent with the hypothesis that the length effect differences between appositives and restrictive relative clauses are due to the at-issueness of the two clause types. Our findings suggest, instead, that implicit prosody plays a role in the length effect observations. We conducted seven experiments manipulating the discourse status and prosody of sen-
tences containing medial and final appositive relative clauses and restrictive relative clauses. Experiment 1 replicated the length effects found in Dillon et al. (2014), in which lengthening a restrictive relative clause led to a greater decrease in the ratings of the containing sentence than lengthening a comparable appositive. Experiments 2 and 3 directly tested whether the at-issueness of appositives could be shown to affect their length penalties. To do this we created a new experimental design that manipulated the at-issue status of appositive content by changing the experimental QUD. This manipulation allowed us to directly test whether it was the at-issue status of appositives and restrictive relatives driving the observed length effects. We provide evidence that, counter to the predictions of a model in which not-at-issue and at-issue content draw from separate working memory resources, acceptability ratings showed the same sensitivity to appositive length whether the appositive contributed not-at-issue or at-issue content. Because we failed to find any effects of at-issueness in Experiments 2 and 3, Experiments 4-7 investigated a prosodic explanation for the length effects using two manipulations of implicit prosody.

**Experiment 1**

Experiment 1 tested the idea that lengthening a restrictive relative clause decreases the acceptability of its containing sentence more than lengthening a comparable appositive clause. To do this, we attempted to closely replicate the first experiment of Dillon et al. (2014); however, we used appositive relative clauses instead of nominal appositives, altered punctuation, and different fillers. The results of experiment 1 showed that medial appositive relative clauses did not add to the perceived complexity of their containing sentences in the same manner as restrictive relative clauses did. These results replicated those of Dillon et al. (2014).

**Method**

**Participants**

Twenty-six participants were recruited from the undergraduate population at the University of California, Santa Cruz, and completed the task in the Linguistics Department Lab. Two participants’ data were excluded because the participants self-reported as non-native English speakers. Participants were instructed to read the sentences at a comfortable speed; completion of the survey took on average 20-30 minutes. All participants received course credit for their participation.

**Materials**

Materials consisted of 24 experimental items comprising the four conditions shown in Table 1. The design was 2 x 2 with the factors length (Long or Short) and clause (Parenthetical or Restrictive). Items were taken from Dillon et al. (2014) and modified in the following ways. Appositive clauses were altered from medial nominal appositive clauses to medial appositive relative clauses (ARCs). Parentheses were also used instead of commas to separate the parenthetical content. An example item is given in (9).

(9) **Dillon et al. (2014) Short Parenthetical Item:**
That man, the one on the cruise, tried to throw a waitress overboard.

**Experiment 1 Short Parenthetical Item:**
The man (who was on the cruise) tried to throw a waitress overboard.

A pilot study of nine participants used em dashes to differentiate the parentheticals from the restrictive relatives, but feedback from participants revealed that this choice of punctuation was confusing. The punctuation was therefore changed in the final experiment to parentheses. Commas were not used because of the concern that they would be an inadequate cue for participants to use to differentiate between parentheticals and restrictive relatives. An example of an experiment item is given below in Table 1.
Table 1 Example of item used in Experiment 1.

Parenthetical, Short
1a. The man (who was on the cruise) tried to throw a waitress overboard.
Parenthetical, Long
1b. The man (who was on the cruise Mary took to the Pacific Islands) tried to throw a waitress overboard.
Restrictive, Short
1c. The man who was on the cruise tried to throw a waitress overboard.
Restrictive, Long
1d. The man who was on the cruise Mary took to the Pacific Islands tried to throw a waitress overboard.

Procedure

The experiment was administered using Linger (http://tedlab.mit.edu/~dr/Linger/). The critical sentences were distributed into four Latin Square lists and were presented in pseudorandom order such that no two critical experimental items appeared adjacent to one another. The 24 experimental items were combined with 72 filler sentences. Fillers were modeled after the description given in Dillon et al. (2014), and comprised a range of grammatical and ungrammatical sentences of similar complexity balanced slightly towards the grammatical end of the range. Of the 72 filler sentences, 40 were judged by three native speaker consultants (who were not included in the actual experiment) on a scale of 1-3 (1 being perfectly grammatical, 3 being ungrammatical, and 2 being in between) as 1, 21 were judged as 3, and 11 were judged as 2. Ungrammatical sentences consisted of phi-feature clashes, island violations, illicit agreement attraction, and resumption violations. In order to balance the appearance of parentheses in the overall items, 12 fillers contained parentheses. These sentences consisted of subject and object parentheticals or nominal appositive clauses.

Participants were instructed to rate the sentences on a 1-7 Likert scale according to how natural each sentence was, 1 being 'very unnatural' and 7 being 'very natural'. Participants were given instructions that included three example sentences and three practice sentences illustrating a range of naturalness scores. Each trial consisted of a single sentence presented on the screen. Participants were instructed to press any key to continue to the ratings page, on which they responded by using either the mouse or the number keys on the keyboard. The Likert scale was labelled on either end of the scale with 'very unnatural' and 'very natural' to remind participants of the direction of the naturalness scale.

Analysis

Results were analyzed using a Cumulative Link Mixed Effects Model with Laplace approximation. A full random effects structure was used.

Results

Overall, we found that adding material to a relative clause decreased the acceptability of the sentence, but there was an interaction such that this decline was greater for a restrictive relative clause than for an appositive relative clause. We found an interaction between length and clause, demonstrating that adding material to a restrictive relative clause decreased the acceptability of the sentence more than adding the equivalent amount of material to a comparable appositive clause. This effect was reflected in an interaction between the length and clause factors, as well as in main effects of length and clause (all ps < .001). Longer sentences received overall lower ratings than shorter sentences, and sentences containing restrictive relatives received overall lower ratings than sentences containing appositives. We computed pairwise comparisons between short and long conditions and found no difference between the short parenthetical
and short restrictive sentences. In the predicted interaction, the rating difference between the short and long restrictive relative clause sentences was greater than the difference between the short and long appositive sentences. The average rating for each critical condition in Experiment 1 is presented in Table 2 and in Figure 1, along with standard errors and length penalties.

<table>
<thead>
<tr>
<th>Clause</th>
<th>Length</th>
<th>Mean</th>
<th>Length Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parenthetical</td>
<td>Short</td>
<td>6.2 (± 0.1)</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>5.8 (± 0.1)</td>
<td></td>
</tr>
<tr>
<td>Restrictive</td>
<td>Short</td>
<td>6.1 (± 0.1)</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>4.7 (± 0.2)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2** Experiment 1 mean acceptability ratings with standard error.

**Discussion**

The results of Experiment 1 affirm the robustness of the length effects found in Dillon et al. (2014). Experiment 1 replicated the findings of Dillon et al. (2014); however, we did so using modified items, a different population, and different fillers. This establishes that the effect is generalizable across appositive clause types, and is not sensitive to the choice of punctuation used to delineate the parenthetical clauses. Under the assumption that ratings data reflect the overall acceptability of a sentence, it follows from the results of Experiment 1 that lengthening a restrictive relative clause led to greater unacceptability of the containing sentence than lengthening a corresponding appositive clause. Assuming additionally that sentence acceptability reflects processing difficulty (Frazier & Clifton, 1989; Sprouse, Fukuda, Ono, & Kluender, 2011; Warren & Gibson, 2002; a.o.), the results suggest that lengthening restrictive relatives burdened the processing domain of matrix sentences more than lengthening appositive clauses did.

Now that we have established the robustness of the length effect, the next two experiments directly test whether we can find evidence that the length effects are due to the different at-issueness status of appositives and...
restrictive relatives. Experiments 2 and 3 test medial and final appositives in a new experimental format that allowed us to manipulate the experimental QUD, and therefore the discourse status of the appositive clauses.

Experiments 2 & 3

Experiments 2 & 3 manipulated the at-issueness of the clauses by altering an explicit experimental QUD. This manipulation tested the claim that the length effects found in Experiment 1 were due to the differing at-issueness of appositives and restrictive relatives. The at-issueness of appositives was manipulated by varying whether the appositive content addressed an explicit QUD in the experimental discourse. Recall that not-at-issue content can be defined as information that does not contribute to resolving the current QUD. As mentioned earlier, there is also a definition of not-at-issue content in which relevance to a QUD does not play a role in determining at-issue status. We do not rule this interesting possibility out, and it is discussed further in the Appendix. However, because relevance to a QUD is an established criterion for at-issueness, we rely on this interpretation here while acknowledging that there is disagreement in the literature. Therefore, we manipulated the at-issueness of the clause by controlling whether the appositive content responds to a QUD in the experimental setting (see also Koev, 2013). On the assumption that the not-at-issue status of appositives is what is driving the length effects, then we predict greater length effects when the appositives are forced to be at issue. Experiment 2 tested medial appositives and Experiment 3 tested final appositives. Both positions were tested because final appositives have been argued to more easily achieve at-issue status (Koev, 2013; Syrett & Koev, 2015).

Method

Participants

Forty workers on Mechanical Turk participated in each experiment; all self-identified as native English speakers and all IP addresses were restricted to the United States.

Materials

The design for each experiment was 2 x 2 x 2, with the factors sentence length (Long and Short), clause (Parenthetical or Restrictive), and at-issueness of the parenthetical content (At-issue or Not-at-issue). Each target sentence was embedded in a multi-exchange discourse in which target sentences were presented as answers to an explicit question in the discourse: at-issue appositives were presented as answers to a coordinated-question QUD, and not-at-issue appositives were presented in answers to a single-question QUD.

Experiment 2 used 32 item sets containing medial appositives and comprised the four conditions given in Table 3. Each sentence was presented under a not-at-issue QUD for the not-at-issue condition, and under an at-issue QUD, for the at-issue condition. Because of the differing definiteness conditions associated with restrictive relative clauses and appositives, materials for Experiment 2 were distributed under two different experimental contexts. The items containing restrictive relative clauses used a context in which there were three of the same type of animal, while the items containing appositives used a context in which there were three animals of different types. Each target condition contained a unique set of animals, which were chosen randomly to exhibit variation in name, kind, and likely familiarity to the participants. Parenthetical and restrictive relative contexts are given in Table 4.
Table 3 Example of item used in Experiment 2.

Not-at-issue QUD: What is the bear wearing?
At-issue QUD: Where is the bear standing and what is it wearing?

Parenthetical, Short
The bear (who is standing on the ball) is wearing a hat.

Parenthetical, Long
The bear (who is standing on the ball the trainer rolled across the room) is wearing a hat.

Restrictive, Short
The bear who is standing on the ball is wearing a hat.

Restrictive, Long
The bear who is standing on the ball the trainer rolled across the room is wearing a hat.

Table 4 Example of Parenthetical and Restrictive QUDs used in Experiment 2.

<table>
<thead>
<tr>
<th>Parenthetical</th>
<th>Context QUD</th>
<th>Viewer Response</th>
<th>At-issue QUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context QUD</td>
<td>What animals do you see in the video?</td>
<td>I see a bear, a cat, and a dog</td>
<td>Where is the bear standing and what is it wearing?</td>
</tr>
<tr>
<td>Viewer Response</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At-issue QUD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restrictive</th>
<th>Context QUD</th>
<th>Viewer Response</th>
<th>At-issue QUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context QUD</td>
<td>What animals do you see in the video?</td>
<td>I see three bears.</td>
<td>Where is one of the bears standing and what is it wearing?</td>
</tr>
<tr>
<td>Viewer Response</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At-issue QUD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experiment 3 used the same set-up as Experiment 2; however, final appositive clauses were used instead of medi-
al clauses as in Experiment 2. An example item is given in Table 5.

Table 5 Example of item used in Experiment 3.

Not-at-issue QUD: What is the giraffe holding?
At-issue QUD: What is the giraffe holding and what is it made of?

Parenthetical, Short
The giraffe is holding a teacup (which is made of a beautiful thin porcelain).

Parenthetical, Long
The giraffe is holding a teacup (which is made of a beautiful thin porcelain the host will easily break).

Restrictive, Short
The giraffe is holding a teacup that is made of a beautiful thin porcelain.

Restrictive, Long
The giraffe is holding a teacup that is made of a beautiful thin porcelain the host will easily break.
Procedure

Critical sentences were distributed into eight Latin Square lists. The 32 critical experimental items from each experiment were combined with 64 filler sentences. In order to ensure that the critical items were not uniquely identifiable by any characteristic, 16 fillers distributed among levels of acceptability used parentheses. The target acceptability distribution of the fillers was 35 as perfectly grammatical, 18 ungrammatical, and 11 in-between. Ungrammatical fillers were created by cycling through the following characteristics: word order mismatches, agreement mismatches, case mismatches, verb selection mismatches, and polarity mismatches. The in-between fillers cycled through the following characteristics: clefts, inversion, inverted pseudo-clefts, and pseudo-clefts. Grammatical fillers cycled through the following structures: reduced relative clauses, coordination, adjunct inclusion, and expletive pronoun constructions (“it appears/seems/looks like...”). The same context QUD was used in the fillers as in the experimental items; in order to balance the single and coordinated-question QUDs in the fillers, half of the fillers contained a single QUD and half contained a coordinated-question QUD.

The experiment was administered on Mechanic Turk and used the IbexFarm experimental software (http://spellout.net/ibexfarm). Participants were told that they would read dialogues involving two people—a Viewer and a Questioner—playing a game over text message. In the game, the Viewer described scenes in a video that they could see but that the Questioner could not see. All the scenes involved a fantasy world in which animals had human-like properties and performed human activities. In each text dialogue, the Questioner asked the Viewer questions about a scene in order to learn what that scene shows. For example, the Questioner asks “What animals do you see in the video?” And the Viewer responds “There are two animals in the video.” Discourses were presented visually as a simulated text message exchange, as shown in Figure 2. Participants were asked to rate how natural the Viewer’s responses to the Questioner were by choosing how natural the response was on a Likert scale of very unnatural (1) to very natural (7). To remind participants of the scale, the Likert scale was labeled on each end as ‘very natural’ and ‘very unnatural’. Three practice items of varying levels of acceptability were given to familiarize participants with the task, and participants were given as long as they liked to judge each item.
**Analysis**

Results were analyzed using a Cumulative Link Mixed Effects Model with Laplace approximation. Experiment 2 used the maximal model that successfully converged (following Barr et al. 2013), which consisted of random intercept for Items and full random slopes for Items. Experiment 3 used a full random effects structure.

**Results**

We predicted that controlling the at-issueness of the appositive clause would cause a greater length effect in the at-issue appositive than in the not-at-issue appositive. Counter to this prediction, we found the same overall interaction between length and clause as found in Experiment 1. This again demonstrates that adding material to a restrictive relative clause decreased the acceptability of the sentence more than adding the equivalent amount of material to a comparable parenthetical clause, regardless of the at-issueness of the clause. This effect was reflected in an interaction between the length and clause conditions, as well as in main effects of length and clause (all $p$s < .001). Longer sentences received overall lower ratings than shorter sentences, and sentences containing restrictive relatives received overall lower ratings than sentences containing appositives. We also found a main effect of at-issueness ($p < .01$), such that the at-issue conditions were rated more highly than the not-at-issue conditions. No other effects reached significance. The average rating for each critical condition in Experiment 2 is presented in Table 6 and in Figure 3, along with standard errors and length penalties.

We additionally analyzed the parenthetical conditions of Experiment 2 separately from the restrictive conditions. We found a significant effect of length ($p < .001$), but no significant difference in means between the at-issue and not-at-issue parentheticals ($p = .084$), and no significant interaction of length and at-issueness ($p = .448$).

Our prediction regarding the at-issueness of appositive clauses was also not born out in final appositives. We found the same interaction between length and clause as found in Experiments 1 and 2. This again demonstrates that adding material to a restrictive relative clause decreased the acceptability of the sentence greater than adding the equivalent amount of material to a comparable parenthetical clause, regardless of the at-issueness of the clause. This effect was reflected in an interaction between the length and clause conditions, as well as in a main effect of length (both $p$s < .001) in which longer sentences received overall lower ratings than shorter sentences. We also found a main effect of at-issueness and an interaction of clause and at-issueness (both $p$s < .05), in
which restrictive relative clauses were rated lower in the not-at-issue condition than in the at-issue condition ($p < .05$). No other effects reached significance. The average rating for each critical condition for Experiments 3 is presented in Figure 4. The average rating for each critical condition for Experiments 2 and 3, along with standard errors and length penalties, is presented in Table 6.

![Figure 4](image_url)

**Figure 4** Mean acceptability ratings of Experiment 3, with final parentheticals and restrictive relative clauses. Error bars indicate standard error of the mean.

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Experiment 2 &amp; 3 mean acceptability ratings with standard error and length penalties.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clause</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td>Parenthetical</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Restrictive</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 3</strong></td>
<td>Parenthetical</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Restrictive</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of Experiment 3 differed from Experiment 2 in that Experiment 3 found no main effect of clause, indicating that the parenthetical sentences were not overall rated higher than the restrictive relative clause sentences. Experiment 3 did find a main effect of at-issueness, which was not found in Experiment 2. This result reflects the finding that restrictive relative clauses were rated lower in the not-at-issue condition at a statistically significant rate, though numerically this effect was small. We can tentatively conclude from this that there was a slight preference for the restrictive relative clauses to carry material
relevant to the QUD, as opposed to merely carrying information that establishes the referent of the head noun phrase.

Discussion

Experiments 2 and 3 tested whether changing the at-issue status of appositives modulated the length effects found in Experiment 1. By providing an explicit QUD in the experimental discourse, we were able to directly manipulate the at-issue status of the experimental appositives, thereby testing both at-issue and not-at-issue appositive clauses. Recall our prediction: if the observed length effects were due to the not-at-issueness of the appositive clauses, then an at-issue appositive should contribute to length effects in the same manner as the at-issue restrictive relative clauses. Our findings in Experiments 2 and 3 did not substantiate this prediction. While we found the same length effects as in Experiment 1—lengthening a restrictive relative caused a greater decrease in acceptability of the containing sentence than increasing a corresponding parenthetical—we observed no significant difference in length effects between the at-issue and not-at-issue conditions. That is, the length effects in the at-issue conditions patterned with the length effects in the not-at-issue conditions. This finding is problematic for a theory in which the length effect difference between parentheticals and restrictive relatives is attributed to their at-issue status.

We also found in Experiments 2 and 3 that the length effect was attenuated in final appositives and relative clauses as compared to medial appositives and relative clauses. This finding was driven largely by the final, long restrictive relatives being rated overall high than the medial, long restrictive relatives, leading to a smaller length penalty for restrictive relative clauses in Experiment 3 than in Experiment 2. Because this finding held across the not-at-issue and at-issue conditions, it is unlikely that we will find an explanation in the at-issueness manipulation. Instead, it may be that the explanation lies in the prosodic domain, and we return to this possibility in the General Discussion.

Experiments 2 and 3 directly tested the hypothesis that the at-issueness difference between restrictive relative clauses and appositives was the main driver behind the length effect difference found in Experiment 1. Our findings suggest that this is not the case, and our experiments revealed no significant modulation of the length effects when the at-issueness of appositives was manipulated. Because we failed to find a significant effect of at-issueness, Experiments 4 and 5 tested the alternative hypothesis that the length effect is due to prosodic differences between parentheticals and restrictive relative clauses.

Experiments 4 & 5

Experiments 4 and 5 tested whether the length effect difference found between appositives and restrictive relative clauses is due to prosodic differences between the clause types. We controlled for the implicit prosody of appositives and restrictive relatives by using adverbial modifiers to force a prosodic break in the critical items. If the degradation of ratings of long restrictive relatives in relation to long parentheticals is due to the difficulty of prosodically parsing the long restrictive relatives, we hypothesized that adding in a forced prosodic break would aid in the parsing of the sentences. Experiment 4 tested medial appositives/restrictive relatives and Experiment 5 tested final appositives/restrictive relatives.

Method

Subjects
Thirty workers on Mechanical Turk\(^2\) and ten participants on TurkPrime participated in Experiment 4,\(^3\) and 44 workers on TurkPrime participated in Experiment 5. All participants self-identified as native English speakers

---

\(^2\) Eight participants were found to have participated in previous experiments in this project. Two analyses were run, one including the duplicate participants and one excluding them. We found no significant differences, so the duplicate participants were included in the final analysis.

\(^3\) The additional ten participants were included to balance the latin square list distribution.
and all IP addresses were restricted to the United States.

**Materials**

The design was 2 x 2 x 2, with the factors sentence length (Long and Short), clause (Parenthetical or Restrictive), and adverb (Adverb or No Adverb). Critical items were those used in Experiments 2 and 3, with an added adverb. In Experiment 4, an adverb was added after the copula in each critical item in the adverb condition. An example item is given in Table 7.

**Table 7** Example item used in Experiment 4.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Adverb Location</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parenthetical, Short, adverb</td>
<td>The bear (who is standing on the ball) is, predictably, wearing a hat.</td>
<td></td>
</tr>
<tr>
<td>Parenthetical, Long, adverb</td>
<td>The bear (who is standing on the ball the trainer rolled across the room) is, predictably, wearing a hat.</td>
<td></td>
</tr>
<tr>
<td>Restrictive, Short, adverb</td>
<td>The bear who is standing on the ball is, predictably, wearing a hat.</td>
<td></td>
</tr>
<tr>
<td>Restrictive, Long, adverb</td>
<td>The bear who is standing on the ball the trainer rolled across the room is, predictably, wearing a hat.</td>
<td></td>
</tr>
</tbody>
</table>

We placed the adverb after the copula so that it created a prosodic break after the restrictive relative clause; the placement of the adverb before the copula was judged to be less felicitous than placing it afterward. In Experiment 5, we placed the adverb after the complementizer in each critical item in the adverb condition. An example item is given in Table 8.

**Table 8** Example item used in Experiment 5.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Adverb Location</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parenthetical, Short, Adverb</td>
<td>The bear is wearing a beret (which, predictably, is a light blue color).</td>
<td></td>
</tr>
<tr>
<td>Parenthetical, Long, Adverb</td>
<td>The bear is wearing a beret (which, predictably, is a light blue color the French trainer picked out).</td>
<td></td>
</tr>
<tr>
<td>Restrictive, Short, Adverb</td>
<td>The bear is wearing a beret that, predictably, is a light blue color.</td>
<td></td>
</tr>
<tr>
<td>Restrictive, Long, Adverb</td>
<td>The bear is wearing a beret that, predictably, is a light blue color the French trainer picked out.</td>
<td></td>
</tr>
</tbody>
</table>

The following adverbs were used for both experiments: amusingly, annoyingly, apparently, astoundingly, bizarrely, clearly, comically, curiously, evidently, hilariously, I assume, I see, I think, I'm sure, improbably, in fact, interestingly, it appears, it looks like, it seems, most likely, oddly, predictably, probably, seemingly, strangely enough, stupidly, surprisingly, unexpectedly.

**Procedure**

The experimental procedure was identical to Experiments 2 and 3. Filler sentences were also identical, except that adverbs were included in 16 fillers in order to mask the adverb variable in the critical items.
Analysis

Results were analyzed using a Cumulative Link Mixed Effects Model with Laplace approximation. Experiment 4 was analyzed using a full random effects structure. Experiment 5 used the maximal model that successfully converged, which consisted of random intercept for Item, Item random slopes for the effects of length and clause, but not their interaction, and a full random slope structure for Participant.

Results

We predicted that using an adverb to force prosodic breaks within the sentences would result in a smaller length penalty for long sentences containing restrictive relative clauses. Counter to this prediction, we found the same interaction between length and clause as found in the previous experiments. This again demonstrates that increasing the length of a restrictive relative clause decreased the acceptability of the sentence more than adding the equivalent amount of material to a comparable parenthetical clause. This effect was reflected in main effects of length and clause, and an interaction of length and clause (all $p < .001$). We also found that sentences in the adverb condition were rated overall lower than in the no adverb condition, and that the adverbs lowered the ratings of the short sentences more than they did the long sentences. This was reflected in a main effect of adverb and an interaction of adverb and length (all $p < .001$). No other effects reached significance. The average rating for each critical condition in Experiment 4 is presented in Table 9 and in Figure 5 below, along with standard errors, length penalties, and effect sizes calculated using Cohen’s d.

<table>
<thead>
<tr>
<th>Clause</th>
<th>Length</th>
<th>Adverb Mean</th>
<th>Length Penalty</th>
<th>Effect Size</th>
<th>No Adverb Mean</th>
<th>Length Penalty</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parenthetical</td>
<td>Short</td>
<td>5.0 (± .1)</td>
<td>.6</td>
<td>0.84</td>
<td>5.9 (± .1)</td>
<td>4.9 (± .1)</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>4.4 (± .1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restrictive</td>
<td>Short</td>
<td>5.1 (± .1)</td>
<td>1.4</td>
<td>1.71</td>
<td>5.9 (± .1)</td>
<td>3.8 (± .1)</td>
<td>2.99</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>3.7 (± .2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference of Difference Scores:</td>
<td>- .8</td>
<td>-1.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9 Mean acceptability ratings, length penalties, and effect sizes for Experiment 4.

Fig. 5 Experiment 4 mean acceptability ratings with standard error for medial appositive and restrictive relatives.
A between-adverb condition analysis was also run. In the adverb condition, we found main effects of clause \( (p < .001) \) and length \( (p < .05) \), and an interaction of length and clause \( (p < .01) \). In the no adverb condition, we found main effects of length and clause, and an interaction effect of length and clause \( (all \, ps < .001) \).

We made the same prediction for Experiment 5 as for 4: using an adverb to force prosodic breaks within the sentences would result in a lesser length penalty for long sentences containing restrictive relative clauses. Again this prediction was not borne out. We found an interaction between length and clause as found in the previous experiments, demonstrating that adding material to a restrictive relative clause decreased the acceptability of the sentence greater than adding the equivalent amount of material to a comparable parenthetical clause. This effect was reflected in a main effect of clause, and an interaction of length and clause \( (all \, ps < .001) \). However, this interaction was largely driven by the no adverb condition, as reflected in a marginally significant interaction of length, clause, and adverb \( (p = .07) \). We also found that, as in Experiment 4, sentences in the adverb condition were again rated overall lower than in the no adverb condition, and that the adverbs lowered the ratings of the short sentences more than they did the long sentences. This was reflected in a main effect of adverb and an interaction of adverb and length \( (all \, ps < .001) \). No other effects reached significance. The average rating for each critical condition in Experiment 5 is presented in Table 10 and in Figure 6 below, along with standard errors, length penalties, and effect sizes calculated using Cohen’s d.

### Table 10
Mean acceptability ratings, length penalties, and effect sizes for Experiment 5.

<table>
<thead>
<tr>
<th>Clause</th>
<th>Length</th>
<th>Adverb Mean</th>
<th>Length Penalty</th>
<th>Effect Size</th>
<th>No Adverb Mean</th>
<th>Length Penalty</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parenthetical</td>
<td>Short</td>
<td>5.4 (± .1)</td>
<td>.6</td>
<td>.88</td>
<td>5.8 (± .1)</td>
<td>.8</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>4.8 (± .1)</td>
<td>5.0 (± .1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restrictive</td>
<td>Short</td>
<td>5.3 (± .1)</td>
<td>.7</td>
<td>.91</td>
<td>6.0 (± .1)</td>
<td>4.6 (± .2)</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>4.6 (± .1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Difference of Difference Score: \(-.1\) \(-.6\)

A between-adverb condition analysis was also run. In the adverb condition, we found a main effect of length \( (p < .001) \). In the no adverb condition, we found a main effect of length \( (p < .001) \) and an interaction of length and clause \( (p < .05) \). No other effects reached significance threshold of \( p < .05 \).
Discussion

Experiments 4 and 5 tested whether manipulating prosodic boundaries would affect the persistent length effects found in Experiments 1-3. Counter to our predictions, we found significant length by clause interactions in both the no adverb and the adverb conditions in Experiment 4. As shown in Table 9, we did find numerically smaller length penalties for both the parenthetical and restrictive relative conditions in the adverb condition as compared to the no adverb condition. However, the smaller length penalty in the adverb condition is not driven by the long restrictive relative receiving higher acceptability ratings. Instead, the short restrictive relative and both parenthetical conditions decreased in acceptability in the adverb condition, while the long restrictive relative condition stayed steady between the adverb and no adverb conditions.

Why was the long restrictive relative not penalized in the adverb condition? We might expect to observe an additive effect of the adverb manipulation, in which the penalty for adding in the adverb should be apparent in all the conditions in which the adverb manipulation was present. One possibility is that the long restrictive relatives are at a ratings floor; this would mean that people are simply not willing to rate grammatical sentences any lower. In this case, any additive effects that might decrease the rating while not creating outright ungrammaticality would not be observed in the rating outcome. A look at our filler ratings for Experiment 4, though, suggests that a ratings floor is unlikely to be the culprit. The average rating for fillers by grammaticality is presented in Table 11, along with standard deviations.

<table>
<thead>
<tr>
<th>Table 11 Experiment 4 Filler Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filler Type</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Grammatical</td>
</tr>
<tr>
<td>In Between</td>
</tr>
<tr>
<td>Ungrammatical</td>
</tr>
</tbody>
</table>

The In-Between fillers were rated at an average of 3.2, while the long restrictive relatives were rated as a 3.7 in the adverb condition and 3.8 in the no adverb condition. Because the In-Between fillers were grammatical sentences, we find evidence that people were in fact willing to rate grammatical sentences below the Long Restrictive means.
We propose that the rating differences were instead driven by the implicit prosody imposed by the adverbs. In short conditions such as (10), the addition of the adverb forces artificially short prosodic phrases, most notably by forcing the copula into its own prosodic phrase. The prosodic ill-formedness of such short prosodic phrases could be enough to drive the rating decreases in the short conditions.

(10)  The bear who is standing on the ball is, predictably, wearing a hat.

In the long parenthetical conditions, such as (11), the adverb again forces a prosodic break in the material following the close parenthesis, which is material that would otherwise likely be parsed into a single prosodic phrase.

(11)  The bear (who is standing on the ball the trainer rolled across the room) is, predictably, wearing a hat.

However, the sentences in the long restrictive condition lack the prosodic cues present in the parenthetical conditions, i.e. they lack parentheses. The sentences are also long enough that creating an initial prosodic parse without any cues is likely challenging for most readers. The addition of the adverb as a prosodic cue may add a benefit that outweighs, or cancels out, the cost of creating a smaller than ideal prosodic phrase.

(12)  The bear who is standing on the ball the trainer rolled across the room is, predictably, wearing a hat.

Experiment 5 showed a significant length by clause interaction in the no adverb condition; however, this interaction did not reach significance in the adverb condition. As in Experiment 4, the decrease in the length penalty in the adverb condition was not driven by an increase in ratings of the long restrictive condition. Instead, the short conditions and the long parenthetical condition again showed a decrease in acceptability compared to the no adverb condition. It is once again unlikely that the long restrictive conditions are merely at a ratings floor, given that they were rated higher than the In-Between fillers from Experiment 5, as shown in Table 12 below.

<table>
<thead>
<tr>
<th>Filler Type</th>
<th>Mean Rating</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical</td>
<td>5.13 (s = 1.59)</td>
<td>1540</td>
</tr>
<tr>
<td>In Between</td>
<td>3.35 (s = 1.87)</td>
<td>572</td>
</tr>
<tr>
<td>Ungrammatical</td>
<td>2.70 (s = 1.85)</td>
<td>704</td>
</tr>
</tbody>
</table>

Overall, we conclude that the adverbs did successfully affect the prosody of the experimental items in both Experiment 4 and Experiment 5. However, the prosodic effects did not decrease the length penalty of the long restrictive relative clauses, as we had hypothesized. While Experiment 5 showed a somewhat attenuated length penalty between appositives and restrictive relative clauses, the attenuation was not driven by increasing the ratings of the long restrictive relative clauses compared to the other conditions. Therefore, while Experiments 4 and 5 showed that manipulating the prosody of the items does change the length penalties, we were not able to successfully eliminate the length effect difference. Because the adverbs generated non-optimal prosodic phrasing, Experiments 6 and 7 use a different methodology to control the implicit prosody of the experimental items.
Experiments 6 & 7

Experiments 6 and 7 drew on methods from Swets et al. (2007), in which line breaks were used to force readers to impose a particular implicit prosody onto the items. Swets et al. presented ambiguous relative clauses in Dutch and English to participants in three timed chunks, delineated here by brackets:

(13) [The servant of the princess] [who scratched herself in public] [is mildly ashamed]

They found that presenting the sentence in chunks instead of all together yielded more N1, or high, attachments in both English and Dutch, which they attribute to the effects of implicit prosodic parsing on syntactic parsing. In Experiments 6 and 7, we used line breaks to impose a particular prosodic parsing onto experimental items without the confound of introducing additional lexical material as in Experiments 4 and 5. Experiment 6 again tested medial appositives, and Experiment 7 tested final appositives.

Methods

Participants
Forty workers on TurkPrime participated in each experiment; all self-identified as native English speakers and all IP addresses were restricted to the United States.

Materials
The design was 2 x 2, with the factors sentence length (Long and Short), and clause (Parenthetical or Restrictive). Thirty-two test items were used. The materials for Experiments 6 and 7 were identical to those in Experiments 4 and 5, respectively, except that the adverbs were removed. All test items were presented in multiple windows: the first window included only the material preceding the parenthetical or restrictive relative clause, the second window included only the parenthetical or restrictive relative clause (including punctuation), and the third window for Experiment 7 included only the material following the parenthetical or restrictive relative clause. An example item for Experiment 6 is given in (14) and (15) below; brackets indicate material in long items and pipes indicate line breaks.

(14) Restrictive
The bear | that is standing on the ball [the trainer rolled across the room] | is wearing a hat.

(15) Parenthetical
The bear | (who is standing on the ball [the trainer rolled across the room]) | is wearing a hat.

Each sentence chunk was presented for a fixed period of time in the center of the screen. Timing was determined by a script that assigned each word a 210ms window, added 75ms for each character, and included a 150ms word pause time. These settings were determined using naïve participants who did not participate in the actual experiment. The timing was set so participants could read at a comfortable speed without lag time at the end.

Sixty-four fillers were included. The fillers were identical to previous experiments, minus the adverbs that were added for Experiments 4 and 5. Fillers were presented in the same manner as the test items. Following Swets et al. (2007), breaks were made only at major syntactic boundaries (see also Selkirk's (1997) alignment constraints). The amount of material in each break varied, so that some windows included a short string of text (one to two words) and others included a long string of text (up to 12 words). This was done to mask the varying length of the text displayed in the test items. Overall, seven fillers were displayed using four windows, 31 were displayed using three windows, and 26 were displayed using two windows.

---

4 Script written by Ethan Poole. Thank you to Brian Dillon and Tom Roberts for their help with its implementation.
Procedure

Critical sentences were distributed into eight Latin Square lists. The experiment was administered on TurkPrime using the IbexFarm experimental software (http://spellout.net/ibexfarm). The procedure was identical to the previous experiments.

Analysis

Results were analyzed using a Cumulative Link Mixed Effects Model with Laplace approximation. Experiment 6 was analyzed using a full random effects structure. Experiment 7 used the maximal model that successfully converged, which consisted of random intercept for Item, and full random slopes for Item and Participant.

Results

We predicted that using line breaks to force a particular prosody onto experimental items would ameliorate the length penalty by increasing the ratings of sentences in the long restrictive relative clause condition. Although we found a quantitatively larger length penalty for restrictive relative clauses than for parentheticals in Experiment 6, the interaction of length and clause did not reach significance ($p = .07$). We therefore found no statistically significant evidence that lengthening the restrictive relative clauses decreased their acceptability more than the parenthetical clauses. As in previous experiments, we found that long sentences were rated overall lower than short sentences, and that restrictive relative clause sentences were rated overall lower than parenthetical sentences. These observations were realized in main effects of length and clause ($ps < .001$). The average rating for each condition in Experiment 6 is presented in Table 13 and in Figure 7, along with standard errors, length penalties, and effect sizes calculated using Cohen's $d$.

![Fig. 7](image)

**Fig. 7** Experiment 6 mean acceptability ratings with standard error for medial appositives and restrictive relative clauses.

As in Experiment 6, we predicted that using line breaks to force a particular prosody onto the fi-
nal restrictive relative clause sentences in Experiment 7 would cause a decrease in the sentences' length penalty. Our results were similar to those found in Experiment 6. Though we again found a quantitatively larger length penalty for restrictive relative clauses, the interaction of length and clause did not reach significance ($p = .08$). We therefore again found no statistically significant evidence that lengthening the restrictive relative clauses decreased their acceptability more than the parenthetical clauses. As in previous experiments we found that long sentences were rated overall lower than short sentences; however, we did not find that restrictive relative clause sentences were rated overall lower than parenthetical sentences. These observations were realized only in a main effect of clause ($p < .001$). The average rating for each critical condition in Experiment 7 is presented in Table 13 and in Figure 8, along with standard errors, length penalties, and effect sizes calculated using Cohen's d.

**Table 13** Experiment 6 & 7 mean acceptability ratings, length penalties, and effect sizes

<table>
<thead>
<tr>
<th>Clause</th>
<th>Length</th>
<th>Mean Rating</th>
<th>Length Penalty</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Short</td>
<td>5.7 (± .1)</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long</td>
<td>4.9 (± .1)</td>
<td></td>
</tr>
<tr>
<td>Parenthetical</td>
<td></td>
<td>Short</td>
<td>5.6 (± .1)</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long</td>
<td>4.2 (± .1)</td>
<td></td>
</tr>
<tr>
<td>Restrictive</td>
<td></td>
<td>Short</td>
<td>5.9 (± .1)</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long</td>
<td>4.4 (± .1)</td>
<td></td>
</tr>
</tbody>
</table>

Difference of Difference Score: -.6

**Fig. 8** Experiment 7 mean acceptability ratings with standard error for final appositives and relative clauses.

**Discussion**

The results of Experiments 6 and 7 suggest that implicit prosody plays a role in the length effect we found in Ex-
periments 1-5. Out of the seven experiments conducted in this paper, Experiments 6 and 7 are the only experiments in which the length effect difference between the parenthetical condition and the restrictive relative condition was neutralized. Although we are arguing from a null result, the robustness of the length effects in Experiments 1-5, along with the lack of a significant result in Experiment 6 and in Experiment 7, together provide strong suggestive evidence that the line break manipulation successfully ameliorated the length effect difference between the two clause types. The failure to find a significant result suggests that prosodic differences between the two clause types is at least a partial determinant of the length effects.

Although the length by clause interaction did not reach significance in either Experiment 6 or 7, we still found an overall effect of long sentences rated lower than short sentences. Indeed, the large effect sizes of the length differences between short and long restrictive relative clause sentences in both experiments suggest that controlling the prosody does not ameliorate the length effect difference within clause type. We also did not find an attenuated length effect in final clauses as compared to medial clauses, as was found in previous experiment setups. This may be because of the nature of the line break manipulation. While the Experiment 6 sentences containing medial clauses used three separate screens, the Experiment 7 sentences containing final clauses used only two screens: one containing the matrix clause, and one containing the final appositive or restrictive relative. It is plausible that the experimental manipulation had less of an effect on the ratings of the sentences in Experiment 7, due to there being only one point in each item at which the manipulation was used.

**General Discussion**

Our findings establish the robustness of the length effect difference between restrictive relative clauses and appositives: the effect has proven stubbornly persistent over multiple experiments and over multiple manipulations. We conclude the paper by proposing a summary of the landscape currently facing us. The results of Dillon et al. and the current experiments suggest that appositives and non-restrictive relative clauses are somehow differentiated in memory, but the exact nature of this separation is still under question. We propose three possible ways to locate the distinction: at the discourse or speech-act level, at the semantic level, or at the prosodic level. Furthermore, there are two possible ways in which the distinction could be made. Dillon et al. proposed that appositives and at-issue content utilize separate working memory stores. We have proposed a related alternative in which parentheticals differentially consume working memory resources, but do not constitute a separate working memory bin from main clause content. Either of these options could hold in principle hold at the discourse level, at the semantic level, or at the prosodic level.

Our findings suggest that it is unlikely that the locus of the length effect is located at the discourse level. Experiments 2 and 3 manipulated the discourse status of the appositive clauses and found no amelioration of the length effect, as would have been expected under an explanation in which appositives and restrictive relatives are differentiated at the speech act level. Instead, the only manipulation under which the length difference was not found was the prosodic line break manipulation in Experiments 6 and 7. The lack of a significant difference between restrictive relatives and appositives in these experiments suggests that the difference may be due to prosodic factors.

Experiments 6-7 provide evidence that encouraging implicit prosodic boundaries within the restrictive relative clause sentences ameliorated the length effect. Our results do not adjudicate between the two possible reasons, discussed earlier, for the prosodic amelioration of the length effect: either prosodic boundaries facilitate parsing, or working memory is sensitive to prosodic phrasing. However, an explanation under which working memory is sensitive to prosodic phrasing is compatible with cognitive architectures in which processing difficulty arises from burdening the working-memory system. For example, a working memory explanation is compatible with the cognitive architecture provided by ACT-R (Lewis &
Vasishth, 2005; Anderson, 2007), in which processing difficulty arises from burdening a system of centrally-communicating buffers and modules containing a highly limited working-memory. A prosodic account under this explanation could propose that incoming appositive material is not necessarily encoded in memory differently than main clause content, but the opportunities to process clausal material at each prosodic break facilitates the processing of incoming material by grouping it into syntactically delineated pieces that can be shuttled through working memory in turns. This explanation, in which working memory is sensitive to prosodic phrasing, allows us to maintain the hypothesis that appositives differentially consume working memory resources from main clause content; however, we identify the locus of the memory resource distinction in prosodic phrasing and not in discourse status. This possibility is consistent with the findings of Hirotani et al. (2006), who found that long and short appositives were read faster than their comparable restrictive relative clauses. They attribute these findings to the implicit prosody imposed by the commas in appositives, which may facilitate integrative processing at that point in the sentence (see also Just & Carpenter, 1980; Hill & Murray, 2000; Warren et al., 2009; Kuperman et al., 2010). If the prosodic and syntactic parsers work in tandem, then we may hypothesize that prosodically grouped constituents, such as appositives, will be less likely to share short term memory resources with their matrix clauses than constituents without such clearly delineated prosodic boundaries, such as restrictive relative clauses.

There is also an alternative interpretation of our results. The first three experiments in this paper directly tested the hypothesis that appositives are differentiated from at-issue content at the level of the discourse or speech-act. As discussed, we found no evidence in favor of this hypothesis. There is, however, an alternative theory of at-issueness. We have treated at-issueness as a variable property that, while characteristic of particular types of constructions (e.g. Potts's conventional implicatures), is not an immutable property of these constructions. This is not the only interpretation of at-issueness in the literature. One can also take a strictly semantic view of at-issueness and claim that conventional implicatures always contribute not-at-issue content, regardless of their discourse status in a particular utterance (Potts 2005). Under this view, appositives will never achieve at-issue status, even if they contain content that responds to a QUD (Potts 2012). However, this view opens up a number of questions. For example, one can wonder how explanatory a notion of at-issueness is if appositives, which can contain primary or secondary discourse information, project beyond propositional attitude verbs or embed under them, and contain new or old information, nevertheless always contribute not-at-issue content. However, it is worth exploring the interpretation of the current findings under such a theory.

A semantic view of appositives, in which they always contribute not-at-issue content, is compatible with the current results. Under such a view, we would not expect the relevance of the appositive content to the resolution of a QUD to change its status as not-at-issue. However, we must be careful when interpreting the processing implications of the findings. Potts (2005) locates the difference between conventional implicatures (note that these are not, even under this strict view, identical to not-at-issue content) and at-issue content in the semantic composition. The motivation for locating the difference in the semantic composition and not at the discourse level comes from observations that conventional implicatures such as appositives act similar to at-issue content for certain discourse-level behavior (i.e. inter-sentential anaphoric dependencies). For example, Potts (2005) observes that appositives act the same as at-issue content in that they can serve as the antecedent of main clause ellipsis and main clause do it/that anaphora. Appositives, like at-issue and unlike presupposed content, can also establish discourse referents. Last, appositives—like at-issue content—can also be used to satisfy presuppositions of subsequent main-clause content. However, Potts argues that, while intersentential anaphoric dependencies hold across at-issue and conventional implicature content, semantic binding dependencies do not. For example, the quantifier in (16) cannot bind into the appositive in (a), while it can bind into the restrictive relative in (b).

\[
\begin{align*}
(16) & \quad a. *\text{No sailor, who tied her; bowline properly, failed her; test.} \\
& \quad b. \text{No sailor, who tied her; bowline properly failed her; test.}
\end{align*}
\]

Therefore Potts concludes that “Overall, [these differences] are an indication that we distinguish the at-issue and CI dimensions in the composition. From the point of view of a discourse as a whole, they are identical” (ibid, pg. 53). In conclusion, we could locate the processing difference between appositives and restrictive relatives at the
semantic composition level and not at the discourse, or speech-act, level. This interpretation appears to run counter to the interpretation given by Dillon et al. (2014), in which the length effect differences are attributed to the processing differentiation of appositive and restrictive relatives because of their status as separate speech acts. However, as this is merely one interpretation given to the results, we believe the discussion here is compatible in general with their conclusions.

There are two mechanisms by which we can instantiate an account under which appositives and restrictive relatives differ at the level of semantic composition. One is the view discussed above: working memory resources are segregated at the level of semantic composition; the difference is therefore structural and is consistent with the current findings that discourse status did not alter the length penalties. The second is a slightly different but related way of instantiating the semantic account, which locates the difference in attention allocation. Because appositive content is semantically independent from main clause content, it may be that we do not spend working memory resources on appositive content because we know that it will not be used to construct main clause dependencies; that is, we simply pay less attention to appositive content. If the effects are due to this type of resource or attentional allocation, then we predict that experiments which force additional attentional allocation onto the appositives will affect the length penalty of the appositive. This explanation was explored in Experiment 2 in Dillon et al. (2014), in which a ungrammaticality detection paradigm was used to assess differences in attention between parentheticals and restrictive relative clauses. Dillon et al. found that agreement errors were not penalized at a statistically significant different rate between parenthetical clauses and restrictive relative clauses. They therefore found no evidence for attentional discrepancies between the two clause types. However, this paradigm probes attention allocation at a very low-level. A follow-up experiment could manipulate attention-allocation at a higher-level by integrating appositives into the matrix clause using anaphoric dependencies across clause boundaries. This manipulation would force interpreters to allocate attention to the parenthetical clause in order to resolve referents and dependencies in the matrix clause. We leave this suggestion to future research.

A second main finding of the experiments here is that the length effect was ameliorated in final appositive and restrictive relative clauses as compared to medial clauses. Figure 9 plots the length penalties for both the medial and final versions of each manipulation. We believe that there are two probable reasons that the length penalty was attenuated in final position. The first is that the prosody of appositives in final position differs from their prosody in medial position. Because final appositives occur at the end of sentences, their closing boundary receives sentence-final prosody, or a boundary tone. This is in contrast to medial appositives, which receive a prosodic contour signaling that the sentence continues after the appositive closing boundary, that is a phrase accent but not a boundary tone or phrase final lengthening (Jun, 2003). If implicit prosody plays a role in the length penalties, then the differing prosodic contours projected onto the clauses in medial and final position may affect the strength of the length penalty. However, because the length effect was attenuated in both parentheticals and restrictives, a prosodic account must explain why final position appears to affect both types of clauses. One possible avenue of testing for this hypothesis is to neutralize the prosodic differences between the medial and final clauses, either by a methodological obfuscation of the prosody of the entire sentences, or by presenting the clauses separately from any final sentence punctuation, thereby eliminating visual cues for generating sentence-final prosody.
The second possibility for the weaker length penalty found in final position is that the clauses in final position did not intervene within a dependency. As discussed earlier, length effects are often attributed to a dis-preference for long dependencies in natural language. In medial position, both appositives and restrictive relative clauses intervened between a subject/verb dependency; however, in final position neither clause intervened within a dependency. Repeated findings of the length effect in final position suggest that the length effect was not due solely to intervention within a dependency formation; however, it is possible that intervening within a dependency created a greater processing cost than when the clauses were merely adding length to the sentence. This explanation is promising in that it attributes the final attenuation to the location of both types of clauses.

Acknowledgments
We thank Pranav Anand, Adrian Brasoveanu, Brian Dillon, Donka Farkas, and audiences at XPrag 2017 and the CUNY Graduate Center’s Psycholinguistics Supper for helpful comments and suggestions.

References