TOWARDS CHARACTERIZING INCREMENTAL STRUCTURE BUILDING DURING SENTENCE COMPREHENSION

by

Grusha Prasad

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

June, 2022

Abstract

Language comprehension involves incrementally processing sequences of words and generating expectations about upcoming words based on prior context. One of the steps involved in incremental processing is *incremental structure building* — i.e., determining the relationship between the words in a sentence as the sentence unfolds. To understand *how* comprehenders build incremental structures, it is necessary to understand *what* structures comprehenders build in the first place and *why*. This dissertation includes three projects that tackle these what and why questions by studying incremental structure building in sentences with reduced relative clauses as a case study. The first project proposes a method for characterizing what incremental structures human comprehenders build. This method involves three steps: first, implement hypotheses from generative syntax about the abstract structure of sentences in a novel computational model; second, use the model to generate quantitative behavioral predictions; and third, test these predictions using a novel web-based experimental paradigm. Applying this approach, we compared two competing theoretical hypotheses about the structure of reduced relative clauses — Whiz-Deletion and Participial-Phrase — and demonstrated that the Whiz-Deletion account better characterizes the incremental structures that human comprehenders build. The second project studies why the incremental structures that comprehenders construct can change depending on the

environment they are in by testing the following widely debated hypothesis: comprehenders maintain probability distributions over the structures they expect to encounter and rapidly update these distributions to match the statistics of their current environment. Based on a large-scaled reading experiment, we find evidence in support of this hypothesis, but also explain why prior work might have failed to find such support. The third project proposes a method for characterizing what incremental structures Artificial Neural Networks build when processing sentences. Applying this method, we demonstrated that the incremental structures these networks build, like the structures built by human comprehenders, is better characterized by the Whiz-Deletion account than the Participial-Phrase account. Thus, by making it possible to compare the incremental structures that these networks build to the structures that humans build, this method in turn makes it possible to test hypotheses about why humans build the structures they do. I propose several directions for future work which involve applying the methods proposed in these projects to study other phenomena beyond reduced relative clauses.

Thesis Committee

Primary Readers

Tal Linzen (Primary Advisor)
Assistant Professor
Department of Linguistics & Center for Data Science
New York University

Géraldine Legendre
Professor
Department of Cognitive Science
Johns Hopkins University

Brian Dillon
Associate Professor
Department of Linguistics
University of Massachusetts, Amherst

Christopher Honey
Assisstant Professor
Psychological and Brain Sciences
Johns Hopkins University

Lisa Feigenson
Professor
Psychological and Brain Sciences
Johns Hopkins University

Alternate Readers

Kyle Rawlins
Associate Professor
Department of Cognitive Science
Johns Hopkins University

Chaz Firestone
Assisstant Professor
Psychological and Brain Sciences
Johns Hopkins University

Acknowledgments

I would like to thank all of the pillars of support in my life without whom this dissertation could not have been in this **Ph**inished**D** state. So a huge thank you to:

Tal Linzen for constantly pushing me to be precise in my thinking and writing while also being an empathetic and practical mentor.

Other members of my committee Géraldine Legendre, Brian Dillon, Christopher Honey and Lisa Feigenson for your time, feedback and a genuinely fun and insightful defense. Thank you also to Géraldine for taking me under your wing and Brian and Chris for welcoming me into your labs at different points in my academic journey.

Other collaborators Marten van Schijndel, Shauli Ravfogel, Yoav Goldberg, Adina Williams, Yixin Nie, Robin Jia, Douwe Kiela, Mohit Bansal, Colin Wilson, Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto and Christian Muxica. I've learnt a lot from the projects we've worked on together.

Other teachers and mentors at Hampshire and Johns Hopkins for introducing me to cool ideas in Cognitive Science and cheering me on from the sidelines. A special shout out to Joanna Morris, Daniel Altshuler, Jane Couperus, Carlos Molina-Vital, Lee Spector, Neil Stillings and Paul Smolensky.

All the friends I have spent time with over the past five years for enriching my life. A special shout out to Andrea, Andy, Ayushi, Brian, Chhavi, Darcy, Donald,

Giulia, Hongru, Jane, Judy, Karl, Kristijan, Kyriaki, Laura, Liliana, Najoung, Natalia, Nicole, Ollie, Rima, Sadhwi, Samhitha, Sebastian, Suhas, Tom and Venkat for all the hang-out sessions, making-fun-of-Grusha moments, walks, venting spaces, and/or fun conversations.

Savio for being my rock, joining me on this roller coaster journey and for helping this Hermione sort out her priorities.

Manasvini for being an inspiring PhD buddy and role-model, reminding me to trust my instincts and for all the not-so-useless advice.

Sai Prasad my dad for our daily routines, for being my stand-by alarm clock and your unconditional love, support and confidence in me.

Aaroha for being the best sister ever, never ceasing to amaze me and for making me feel Jing.

Srimathi and Narayana my ammamma and ajja for being cool and inspiring grand parents.

Sudha for being the most progressive and supportive ajji. You filled my life with love, cuteness and wisdom. I wish you were here to celebrate this moment, but en madlike aagathe?

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sentence comprehension is an extremely complex cognitive task that involves many steps. For example, imagine someone reading the following sequence of words: "Aaroha saw that it was raining outside". In order for the reader to understand the sequence, they have to first recognize the individual letters on the paper (or screen), put them together to identify the words they form, retrieve the meaning of each word in the sequence and then figure out how these individual word meanings combine to generate the intended meaning. Despite the many steps involved, any literate person familiar with English can read the sentence and instantly understand what it means. How are humans able to read and understand sentences so rapidly and effortlessly?

One explanation for rapid and effortless sentence comprehension is that people process words in sentences *incrementally* and generate expectations (or *predictions*) about upcoming words based on the words they have read so far. For example after reading the sequence "Aaroha saw that it was raining outside and took out her ...", the reader might predict that the upcoming word is more likely to be "umbrella" or "raincoat" than "sunglasses". Predicting upcoming words in this manner can help readers process words more quickly if these predictions turn out to be accurate

(Kutas, DeLong, and Smith, 2011). Evidence for incremental predictive processing comes from decades of psycholinguistic experiments which have demonstrated that participants find words that violate their expectations more difficult to process than words that don't — for example, in the previous example, participants would find the word "sunglasses" more difficult to process than the word "umbrella". For a review and discussion of this literature see Kuperberg and Jaeger (2016) and Traxler (2014).

In this dissertation, I propose to study how people's predictions about upcoming words are shaped by the incremental *structures* they construct — i.e., their expectation about how the words they have read so far are related to each other. For example, consider the following sequence of words "The graduate student examined...". Most people reading this sequence will construct an incremental structure in which the verb *examined* describes the main event in the sentence (i.e., an event in which someone examined something); and the subject *the graduate student* refers to the agent of the verb (i.e., the person doing the examining). Based on this incremental structure, most readers will predict words that describe something or someone that the graduate student examined, such as in the sentence below.

(1) The graduate student examined the argument carefully.

The incremental structure described above, which I will refer to as the *Main-clause or MC structure*, is only one of the possible structures that a person reading the sentence could have built — i.e., the sequence is *temporarily ambiguous*. An alternative, equally valid structure is one in which the graduate student was not the agent of the examining event, but rather the patient (i.e., the person being examined). In this structure, the verb examined is inside a *relative clause*, and therefore provides additional information

about *the graduate student* instead of describing the main event. We will refer to this alternative structure as the *Relative-Clause or RC structure*. A reader who constructs an RC structure instead of the MC structure will predict words that describe agents of the examining events like in the sentence below, instead of patients (like in (1)); the square brackets indicate the boundaries of the relative clause.

(2) The graduate student [examined **by the committee**] defended the argument.

While both sentences (1) and (2) are grammatical in English, readers encounter active sentences like (1) much more frequently than passive sentences like (2). Therefore, given a temporarily ambiguous sequence like "the graduate student examined", readers are more likely to predict words that are consistent with a MC structure than with a RC structure. In the rare event when these ambiguous sequences are disambiguated in favor of the RC structure, as in (2), readers are surprised and need to update their expectations. Consequently, they read the disambiguating words in sentences like (2) (bolded) more slowly than the same words in minimally different unambiguous sentences like (3) (MacDonald, Pearlmutter, and Seidenberg, 1994; Trueswell, 1996). Sentence (3) is an example of a *full* relative clause which, unlike the *reduced* relative clause in (2), has a wh-phrase "who" and a finite auxiliary "was"; the presence of these disambiguating words early in the sentence ensures that readers never construct the MC structure in the first place.

(3) The graduate student who was examined **by the committee** defended the argument.

This difference in reading times between words in (2) and (3) is often referred to as a

garden path effect. As discussed above, garden path effects are a result of violated expectations about the structure of an ambiguous sequence of words. Therefore, these effects are interpreted as evidence for the hypothesis that readers construct *incremental structures* based on the words words they have read so far, and use these structures to generate expectations about upcoming words.

There is converging evidence for incremental structure building during sentence comprehension beyond the garden path effects described above: experiments have found distinctive electrophysiological patterns that are associated with a reanalysis process when participants construct incorrect incremental structures (Hagoort, Brown, and Groothusen 1993; Kim and Osterhout 2005; for a review and discussion see Van Petten and Luka 2012); there has been evidence for lingering effects of such incorrect incremental structure construction in both offline measures such as participants' responses to comprehension questions (Ferreira, Christianson, and Hollingworth, 2001; Christianson et al., 2001) or their paraphrases of target sentences (Patson et al., 2009) as well as online measures such as increased reading times (Slattery et al. 2013; for a review and comparison of the online and offline measures see Qian, Garnsey, and Christianson 2018); and finally, experiments using the visual world eye-tracking paradigm have found increased anticipatory looks at specific target objects when these objects are plausible given the preferred incremental parse of the preceding words relative to when they are not (Arai and Keller, 2013; Arai, Van Gompel, and Scheepers, 2007).

Given that language comprehension involves incremental structure building, characterizing *what* structures comprehenders build and *why* is crucial a part of solving the bigger puzzle of *how* people understand sentences. In this dissertation I describe

three projects that tackle these *what* and *why* questions, using the comprehension of sentences with relative clauses (such as (2)) as a case study. The order in which these projects are presented in this dissertation does not correspond to the order in which these projects were undertaken, and therefore the later chapters do not reference or build on the arguments and conclusions of the previous ones. Consequently, each chapter is meant to be comprehensible on its own and does not require the reader to incrementally parse the previous chapters. In the remainder of this chapter, I briefly motivate the questions explored in these projects, describe the methods used to study these questions and summarize the main conclusions. Then, in the final chapter of this dissertation, I synthesize the questions emerging from these projects and propose future work.

Note on pronoun use Since all of the three projects were collaborative, the pronoun "we" is used when describing the content of these projects. However, since the description of the broader context of these projects and the questions that emerge from them was written solely for the purposes of this dissertation, Chapters 1 and 5 use the pronoun "I".

Chapter 2: What is the system of rules that governs the incremental structures that human comprehenders build?

In this chapter we argued that theories from generative syntax — a field that studies what sentence structures can and cannot exist across languages — are a useful starting point for generating hypotheses about the system of rules that shapes the incremental structures that readers build during sentence comprehension (or *grammar*). Adopting such a syntactic-theory first approach can help constrain the hypothesis space of all

possible grammars. For a review of prior psycholinguistic work that adopts such a syntactic-theory first approach and a discussion of when and why such an approach is fruitful, see Kush and Dillon (2021) and Phillips et al. (2021).

To convert representational hypotheses from different syntactic theories into testable behavioral predictions, we implemented these different hypotheses in a new model of parsing we proposed: the model of Serial Parsing in ACT-R With Null-elements (or SPAWN). The structures that this model decides to build at any given point (i.e., its *parsing decisions*) are based on the computational principles proposed by a general purpose cognitive architecture, Adaptive Control of Thought-Rational (ACT-R; Anderson et al. 2004), which is designed to explain cognition across a wide range of tasks and domains. Since SPAWN uses ACT-R principles, there is a transparent link between the parsing mechanism and the cognitive processes thought to be involved in sentence comprehension (such as memory retrieval). This transparent link makes it possible to describe the consequences of different representational assumptions on real-time sentence comprehension.

We used SPAWN to evaluate two hypotheses from generative syntax about the structure of sentences with reduced relative clauses such as (2): the **Whiz-Deletion hypothesis** (Chomsky et al., 1957; Ross, 1967; Smith, 1961) and the **Participial-Phrase hypothesis** (Harwood, 2018; Bhatt, 1999; Kayne, 1994). The Whiz-Deletion hypothesis argues that the structure of all relative clauses, reduced or not, share some common properties which are absent from minimally different sentences without relative clauses. The Participial-Phrase hypothesis, on the other hand, argues that these shared properties are absent from the structures of some reduced relative clauses. We implemented these two hypotheses as two separate grammars and used these

grammars to train separate instances of SPAWN models, which we referred to as the Whiz-Deletion and Participial-Phrase versions. From these two versions, we generated predictions about the extent to which participants are expected to complete ambiguous target prompts (e.g., "the graduate student examined") with completions that are consistent with a reduced RC reading (e.g., sentence (2)) when these target prompts are preceded by sentences with different types of RCs.

The predictions we generated were based on the structural priming paradigm (Branigan et al., 1995; Branigan and Pickering, 2017). In this paradigm, participants are presented with *target* sentences (e.g., sentences with reduced RCs). These target sentences are preceded by different types of *prime* sentences, each of which are hypothesized to share different structure properties with the target sentence (e.g., sentences with different types of RCs). The incremental structures that readers construct are inferred by measuring the extent to which each of the different types of primes facilitate the production or comprehension of the target sentence: if some prime A shares some property P_A with the target that is not present in a minimally different sentence prime B, and if A primes the target more than B, then we can infer that incremental structure that readers (implicitly) built when reading the target sentence contains P_A.

We tested the behavioral predictions generated by the Whiz-Deletion and Participial phrase versions of SPAWN by using a novel web-based comprehension-to-production priming paradigm. The empirical data from our human experiment qualitatively aligned with the data from the Whiz-Deletion version of SPAWN, suggesting that as the Whiz-Deletion account assumed, the structures that our participants constructed when processing sentences with different types of RCs shared some common

properties. This representational assumption was not sufficient to account for all of the patterns in the model and human data: some of the patterns in the models emerged from an interaction between the grammar and the parsing mechanism we assumed, suggesting that a similar interaction might be driving our participants' behaviour. This observation highlights the importance of using an explicit model of parsing, even when the focus of the research program might be to characterize the grammar.

At the same time, the predictions from the Whiz-Deletion version of SPAWN underestimated the *magnitude* of priming observed in the empirical data. We argued that the lack of quantitative alignment between the model predictions and empirical data was a consequence of our simplifying assumption that the model only constructed only one incremental structure at a time (i.e., the parsing mechanism was strictly serial); thus, we argued that in order to fully account for the empirical results, a parallel parsing mechanism was likely required (cf. Boston et al. 2011).

Chapter 3: Are the structures that the system of rules builds in temporarily ambiguous sentences impacted by context-specific probabilities?

Under a rational (Anderson, 1990) account of sentence comprehension, the optimal behavior for a person reading a sentence is to either construct incremental structures consistent with the most probable parse given the words they have read so far (probabilistic serial parsing; e.g., Ambati et al., 2015; Yang and Deng, 2020 and the SPAWN model in Chapter 2)[1] or, alternatively, construct all (or several) possible incremental structures but assign the highest weight to the structure consistent with the most

probable parse (probabilistic parallel parsing; e.g., Hale, 2001; Jurafsky, 1996). Both of these strategies require the reader to maintain a probability distribution over parses which is aligned with the distribution of structures in the environment: for example, given the sequence "the graduate student examined...", an optimal reader will assign a higher probability to the MC parse than the RC parse because sentences disambiguated in favor of a MC parse (like (1)) are more frequent than sentences disambiguated in favor of a RC parse (like (2)).[2]Since such a probability distribution can vary drastically across environments and contexts — for example, complex sentence structures that are frequent in formal writing are less frequent in social media posts — it is necessary for rational readers to rapidly adapt their expectations to match the statistics of their current environment (Fine et al., 2013).

Based on the literature discussed earlier, which demonstrated that humans generate expectations about upcoming words based on prior context, there is general consensus in the field that the human sentence comprehension is rational.[3] Yet, the empirical picture is not as straightforward about whether humans rapidly adapt their expectations to match the statistics of an experimental setting, as would be expected of a rational sentence comprehender. Early work explored this question using the self-paced reading paradigm and found that participants who were repeatedly exposed to sentences with reduced relative clauses such as (2) were less surprised when they

[1]While the most optimal choice would be to always select the parse with the highest probability, the SPAWN models adopt a "probability-matching" approach where parses with low probability are occasionally chosen. This aligns with evidence that suggests that people use probability matching instead of maximizing the probability both when making explicit decisions (e.g., Lo, Marlowe, and Zhang 2021) and with more implicit cognitive processes like perception (e.g., Wozny, Beierholm, and Shams 2010)

[2]Serial parsers which use non-probabilistic strategies, such as the two-stage model proposed by Frazier and Fodor (1978), do not require a probability distribution over parses. However, these parsing mechanisms do not presume a rational account of sentence processing.

[3]Or at rational in a cognitively bounded manner (Lewis and Howes, 2020).

encountered these sentences later on in the experiment compared to participants who were exposed to filler sentences (Fine et al., 2013). Based on this result, the authors concluded that human comprehenders do calibrate their expectations rapidly, i.e., there is evidence for *syntactic adaptation*. However, subsequent work with considerably more participants and items failed to replicate this between-group difference, suggesting that any change in reading times over the course of the experiment was driven by familiarity with the experimental task or *task-adaptation* (Stack, James, and Watson, 2018). This replication failure challenges the predictions of the rational account of sentence comprehension.

The goal of Chapter 3 was to clarify the empirical picture regarding rapid syntactic adaptation in self-paced reading. In a large between-group study (n = 642) with a simpler experimental design than in previous work, we found evidence for syntactic adaptation over and above task adaptation, providing further evidence that the human sentence comprehension is rational. However, this effect of syntactic adaptation was very small relative to that of task adaptation, which explains why Stack, James, and Watson might have failed to find this effect. Post-hoc power analyses indicated that a large number of participants were required to detect syntactic adaptation in future between-subjects self-paced reading studies. This issue is exacerbated in experiments designed to detect modulations of the basic syntactic adaptation effect, such as experiments exploring what properties of the reduced RC sentences participants might be adapting to; these experiments are likely to be under-powered even with more than 1200 participants. We concluded that while syntactic adaptation can be detected using self-paced reading, this paradigm is not very effective for studying this phenomenon.

Chapter 4: What is the system of rules that governs the incremental structures that neural networks build?

In Chapter 2, we proposed a method for testing hypotheses about how the system of rules underlying incremental structure building (or *grammar*) is organized in humans. But why is the grammar organized in one way and not another? One approach to answering this question involves generating hypotheses about the role of specific factors in shaping the grammar's organization, and testing these hypotheses by running targeted experiments that manipulate these factors and measure the resulting change to the grammar. Running such targeted experiments on humans might not always be possible. For example, consider the hypothesis that people's grammar is shaped by memory limitations during childhood (cf. Elman 1993). To test this hypothesis, it is necessary to manipulate childrens' memory as they are acquiring langauge, which is not possible (and not ethical to do even if it were). Given the limits of the kinds of experiments we can run on humans, an alternative approach is to run these experiments on models that closely mimic human's ability to comprehend sentences. Modern Artificial Neural Networks (ANNs) have been remarkably successful on a variety of natural language understanding tasks, making them viable candidates on which to run such targeted experiments. However, unlike with the SPAWN model proposed in Chapter 2, the processes and representations underlying the ANNs' behavior are not transparent. Therefore, in order to study how the organization of the grammar these models (implicitly) implement changes as a function of some factor, it is necessary to characterize how this grammar is organized in the first place. In Chapter 4, we propose a method to do so.

This method is inspired by the structural priming paradigm from psycholinguistics

introduced earlier in this chapter. We argue that we can infer that the grammar implicitly implemented by the model is sensitive to some property P if a target sentence with this property is primed more by a prime sentence that also shares this property than by a minimally different sentence without this property. We measured priming in ANNs by measuring how the probability that the models assign to words in the target sentence changes as a consequence of having processed the prime sentence.

As a case study, we applied our method to study the organization of the part of the models' implicitly implemented grammar that is involved in processing sentences with relative clauses. We demonstrated that, like in our human experiments in Chapter 2, sentences with reduced passive relative clauses were primed more by sentences with full passive relative clauses than by sentences without relative clauses. Therefore, the incremental structures constructed by these neural network models, like those constructed by our participants, are better described by the Whiz-Deletion account compared to the Participial-Phrase account. Therefore by studying what factors shape the grammar of these models, we can draw inferences about why readers' grammars might be organized the way they are.

Below I list two not mutually-exclusive hypotheses that I plan to test in future work about why the Whiz-Deletion account might better describe the structures that humans and ANNs construct.

Hypothesis 1: Grammar organization is shaped by memory limitations The first hypothesis is that the grammar of human comprehenders is shaped by memory limitations, particularly those that exist during early stages of language acquisition (cf. *the starting small hypothesis*; Elman 1993). These memory limitations can result in

learners acquiring relative clauses with a shorter distance between the filler and the gap such as (4) more rapidly than relative clauses with a longer distance such as (5) (Gibson, 1998).

(4) The graduate student who ___ examined the committee members was happy.

(5) The graduate student the committee members examined ___ was happy.

Since all relative clauses share some common properties — for example, the lexical items in the relative clause (i.e., *committee members* in (4) and (5)) are not relevant when computing the agreement features of the main verb (i.e., *was*) — an efficient strategy for relative clause acquisition is to identify these shared properties from the already acquired relative clause types and generalize them to the new types. Such a generalization is more feasible with the representations proposed by the Whiz-Deletion account, in which all relative clauses share some common structure, than with those proposed by Participial-Phrase account, in which rapidly acquired active subject relative clauses like in sentence (4) do not share common structure with passive reduced relative clauses like (6)

(6) The graduate student examined by the committee members was happy.

This hypothesis predicts that adding memory limitations to ANNs during language learning, similar to those in human children, will result in these models organizing their grammars more hierarchically than the grammars of models without these added limitations.

Hypothesis 2: Grammar organization is shaped by statistical properties of the linguistic data An alternative and not mutually exclusive hypothesis is that the grammar is shaped by statistical properties of the linguistic data they have been exposed to (as children or adults). For example, let us again consider the case of relative clauses: only 1% of nouns across three large corpora with English sentences were modified by passive reduced relative clauses such as (6); however, when considering nouns modified by *any* type of relative clause, this percentage was much higher at about 13% (Roland, Dick, and Elman, 2007). Therefore a learner who recognizes that passive reduced relative clauses are a type of a larger category of relative clauses (as suggested by the Whiz-Deletion account), will have more training instances from which they can learn the correct agreement behavior in sentences like (6) than a learner who treats passive reduced relative clauses as a category of its own distinct that is from other relative clauses.

I plan to test these two hypotheses in future work by systematically varying the linguistic input that these models are trained on as well as their memory capabilities at early stages of training and measuring how these factors influence the grammar implicitly implemented by ANNs.

Summary

To summarize, this dissertation consists of three projects which use relative clause comprehension as a case study to tackle the questions of *what* incremental structures are built during sentence comprehension and *why*. In Chapter 3 we studied how the structures that humans construct is influenced by the statistical properties of their current environment. In Chapters 2 and 4 we developed methods to characterize the

system of rules (or *grammar*) underlying the incremental structures that humans and ANNs build during sentence comprehension. These methods make it possible to run targeted experiments on ANNs to gain insight into *why* the system of rules used by human comprehenders might be organized the way it is. In the final chapter of this dissertation, I propose two directions for future work that involve applying the methods developed in Chapter 2 to a wider range of psycholinguistic phenomena.

References

Kutas, Marta, Katherine A DeLong, and Nathaniel J Smith (2011). "A look around at what lies ahead: Prediction and predictability in language processing." In.

Kuperberg, Gina R and T Florian Jaeger (2016). "What do we mean by prediction in language comprehension?" In: *Language, cognition and neuroscience* 31.1, pp. 32–59.

Traxler, Matthew J (2014). "Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing". In: *Trends in cognitive sciences* 18.11, pp. 605–611.

MacDonald, Maryellen C., Neal J. Pearlmutter, and Mark S. Seidenberg (1994). "The lexical nature of syntactic ambiguity resolution." In: *Psychological Review* 101.4, pp. 676–703. URL: http://dx.doi.org/10.1037/0033-295X.101.4.676.

Trueswell, John C. (1996). "The Role of Lexical Frequency in Syntactic Ambiguity Resolution". In: *Journal of Memory and Language* 35.4, pp. 566–585. URL: https://doi.org/10.1006/jmla.1996.0030.

Hagoort, Peter, Colin Brown, and Jolanda Groothusen (1993). "The syntactic positive shift (SPS) as an ERP measure of syntactic processing". In: *Language and cognitive processes* 8.4, pp. 439–483.

Kim, Albert and Lee Osterhout (2005). "The independence of combinatory semantic processing: Evidence from event-related potentials". In: *Journal of memory and language* 52.2, pp. 205–225.

Van Petten, Cyma and Barbara J Luka (2012). "Prediction during language comprehension: Benefits, costs, and ERP components". In: *International Journal of Psychophysiology* 83.2, pp. 176–190.

Ferreira, Fernanda, Kiel Christianson, and Andrew Hollingworth (2001). "Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis". In: *Journal of psycholinguistic research* 30.1, pp. 3–20.

Christianson, Kiel, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira (2001). "Thematic roles assigned along the garden path linger". In: *Cognitive psychology* 42.4, pp. 368–407.

Patson, Nikole D, Emily S Darowski, Nicole Moon, and Fernanda Ferreira (2009). "Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.1, p. 280.

Slattery, TJ, P Sturt, K Christianson, M Yoshida, and F Ferreira (2013). "Lingering misinterpretations of garden path sentences arise from flawed semantic processing". In: *Journal of Memory and Language* 69, pp. 104–20.

Qian, Zhiying, Susan Garnsey, and Kiel Christianson (2018). "A comparison of online and offline measures of good-enough processing in garden-path sentences". In: *Language, Cognition and Neuroscience* 33.2, pp. 227–254.

Arai, Manabu and Frank Keller (2013). "The use of verb-specific information for prediction in sentence processing". In: *Language and Cognitive Processes* 28.4, pp. 525–560.

Arai, Manabu, Roger PG Van Gompel, and Christoph Scheepers (2007). "Priming ditransitive structures in comprehension". In: *Cognitive psychology* 54.3, pp. 218–250.

Kush, Dave and Brian Dillon (2021). "Sentence Processing and Syntactic Theory". In: *A Companion to Chomsky*, pp. 305–324.

Phillips, Colin, Phoebe Gaston, Nick Huang, and Hanna Muller (2021). "Theories all the way down: Remarks on "theoretical" and "experimental" linguistics". In: *The Cambridge handbook of experimental syntax*, pp. 587–616.

Anderson, John R, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin (2004). "An integrated theory of the mind." In: *Psychological review* 111.4, p. 1036.

Chomsky, Noam et al. (1957). "Syntactic Structures". In: *The Hague: Mouton*.

Ross, John Robert (1967). "Constraints on variables in syntax." In.

Smith, Carlota S (1961). "A class of complex modifiers in English". In: *Language* 37.3, pp. 342–365.

Harwood, William (2018). "Reduced Relatives and Extended Phases: A Phase-Based Analysis of the Inflectional Restrictions on English Reduced Relative Clauses". In: *Studia Linguistica* 72.2, pp. 428–471.

Bhatt, Rajesh (1999). "Covert modality in non-finite contexts: University of Pennsylvania dissertation". In.

Kayne, Richard S (1994). *The antisymmetry of syntax*. Vol. 25. MIT press.

Branigan, Holly P, Martin J Pickering, Simon P Liversedge, Andrew J Stewart, and Thomas P Urbach (1995). "Syntactic priming: Investigating the mental representation of language". In: *Journal of Psycholinguistic Research* 24.6, pp. 489–506.

Branigan, Holly P and Martin J Pickering (2017). "An experimental approach to linguistic representation". In: *Behavioral and Brain Sciences* 40.

Boston, Marisa Ferrara, John T Hale, Shravan Vasishth, and Reinhold Kliegl (2011). "Parallel processing and sentence comprehension difficulty". In: *Language and Cognitive Processes* 26.3, pp. 301–349.

Anderson, John R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Ambati, Bharat Ram, Tejaswini Deoskar, Mark Johnson, and Mark Steedman (2015). "An incremental algorithm for transition-based CCG parsing". In: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 53–63.

Yang, Kaiyu and Jia Deng (2020). "Strongly incremental constituency parsing with graph neural networks". In: *Advances in Neural Information Processing Systems* 33, pp. 21687–21698.

Hale, John (2001). "A Probabilistic Earley Parser As a Psycholinguistic Model". In: Pittsburgh, Pennsylvania: Association for Computational Linguistics, pp. 1–8. DOI: 10.3115/1073336.1073357. URL: https://doi.org/10.3115/1073336.1073357.

Jurafsky, Daniel (1996). "A probabilistic model of lexical and syntactic access and disambiguation". In: *Cognitive science* 20.2, pp. 137–194.

Lo, Andrew W, Katherine P Marlowe, and Ruixun Zhang (2021). "To maximize or randomize? An experimental study of probability matching in financial decision making". In: *PloS one* 16.8, e0252540.

Wozny, David R, Ulrik R Beierholm, and Ladan Shams (2010). "Probability matching as a computational strategy used in perception". In: *PLoS computational biology* 6.8, e1000871.

Fine, Alex B., T. Florian Jaeger, Thomas A. Farmer, and Ting Qian (2013). "Rapid Expectation Adaptation during Syntactic Comprehension". In: *PLoS One* 8.10, e77661. URL: https://doi.org/10.1371/journal.pone.0077661.

Frazier, Lyn and Janet Dean Fodor (1978). "The sausage machine: A new two-stage parsing model". In: *Cognition* 6.4, pp. 291–325.

Lewis, Richard L and Andrew Howes (2020). "Cognitively bounded rational analyses and the crucial role of theories of subjective utility". In: *Behavioral and Brain Sciences* 43.

Stack, Caoimhe M. Harrington, Ariel N. James, and Duane G. Watson (2018). "A failure to replicate rapid syntactic adaptation in comprehension". In: *Memory and Cognition* 46.6. DOI: 10.3758/s13421-018-0808-6.

Elman, Jeffrey L (1993). "Learning and development in neural networks: The importance of starting small". In: *Cognition* 48.1, pp. 71–99.

Gibson, Edward (1998). "Linguistic complexity: Locality of syntactic dependencies". In: *Cognition* 68.1, pp. 1–76.

Roland, Douglas, Fredric Dick, and Jeffrey L. Elman (2007). "Frequency of basic English grammatical structures: A corpus analysis". In: *Journal of Memory and Language* 57.3, pp. 348–379. URL: https://doi.org/10.1016/j.jml.2007.03.002.

Chapter 2

What is the system of rules that governs the incremental structures that human comprehenders build?

2.1 Introduction

A crucial step in comprehending a sequence of words involves constructing a structural description which specifies the relationship between the different words in the sequence. One of the goals of psycholinguistics is to understand how humans construct these structural descriptions in real time. To understand this, we need to first generate and test hypotheses about *what* structural descriptions humans are constructing. A promising source for these hypotheses is theories from generative syntax – a field that studies the abstract (i.e., not directly observable) relationship between words in sequences by characterizing what structures can (and cannot) exist across human languages.

In this chapter we propose a method of converting assumptions about abstract structure from theories in generative syntax into concrete testable behavioral predictions, thereby providing a principled way of evaluating the different theoretical

assumptions. As a case study, we apply this method to evaluate two different assumptions about the structure of sentences with passive reduced relative clauses (RRCs) such as (1).

(1) The defendant examined by the lawyer was unreliable. (Passive Reduced RC; RRC)

When comprehenders are incrementally parsing sentences like (1), they can construct two structural representations after reading the partial sequence "the defendant examined": a preferred but eventually incorrect main-verb reading corresponding to the canonical transitive argument structure (the defendant examined someone) and a dispreferred but eventually correct reduced RC reading corresponding to a passive use of the verb *examine* (the defendant was examined by someone). This temporary ambiguity has been widely used in psycholinguistics to study the mechanism underlying human incremental sentence processing (for a review see Frazier (2013), McRae and Matsuki (2013), Levy (2013), and Spivey, Anderson, and Farmer (2013)).

These temporarily ambiguous RRC sentences are an ideal case study for our proposed method for two reasons. First, characterizing what structures comprehenders construct when reading RRC sentences is necessary to answer some of the questions raised by the influential body of work studying these temporarily ambiguous sentences such as what expectations about their environment are participants updating when they read sentences with RRCs (cf. Prasad and Linzen 2021). Second, there are two competing theoretical accounts of RRCs — the Whiz-Deletion account (Chomsky et al., 1957; Ross, 1967) and the Participial-Phrase account (Harwood, 2018) — and therefore, it is unknown which of these accounts better describe the representations

that comprehenders construct.

We start by outlining the differences between the Whiz-Deletion and Participial-Phrase. We do so describing how the structure that these accounts assign to sentences like (1) differs from the structure they assign to minimally different sentences like (2) and (3).

(2) The defendant who was examined by the lawyer was unreliable. (Passive Full RC; FRC)

(3) The defendant being examined by the lawyer was unreliable. (Progressive Passive Reduced RC; ProgRRC)

Then, we use the structural priming paradigm to generate behavioral predictions from these two accounts. The logic underlying this paradigm, as described in Chapter 1, is as follows: for two sentences, a prime and a target, which share some properties, if the target is easier to process or produce when it is preceded by the prime than when it is preceded by a minimally different control sentence without these properties, then it can be inferred that the human processing or production system is sensitive to the properties shared by the two sentences (Branigan et al., 1995; Branigan and Pickering, 2017).[1] Using this logic, we can describe the structural differences between the two accounts in terms of the extent to which target sentences with RRCs are expected to be primed by other sentences with RRCs, FRCs or ProgRRCs. The different patterns of priming predicted by the two accounts can then be evaluated against empirical

[1]Priming can also result in *inhibitory* effects — i.e., if a prime-target pair share some property that the human processing or production system is sensitive to, then the target might be more *difficult* to process when it is preceded by the prime than when it is not (Branigan et al., 1995). While such inhibitory effects are logically possible, most work using the priming paradigm focuses on facilitatory effects. Therefore, in this chapter we will use the word *priming* as being synonymous with facilitation.

priming data.

To generate priming predictions, we develop a model of parsing which we call the model of Serial Parsing in ACT-R With Null Elements (SPAWN) and create two versions of this model, one which uses the grammar specified by the Whiz-Deletion account and another which uses the grammar specified by the Participial-Phrase account (§ 2.3). Then, for each of these versions we estimate the probability with which the model assigns a reduced RC parse to an ambiguous *target* sequence like "the defendant examined" (like in (1)) when the model was previously presented with either other *prime* sentences with RRCs, FRCs or ProgRRCs (§ 2.4); the greater the probability with which the model assigns a reduced RC parse in some priming condition, the greater the priming effect.

Finally, we test these predictions by running a comprehension-to-production priming experiment in which we estimated the probability with which participants completed ambiguous target prompts with a reduced RC continuation in the different priming conditions (§ 2.5). To foreshadow our results, we find that the empirical data better aligns with the predictions from the Whiz-deletion account than with the Participial-Phrase account, suggesting that the Whiz-deletion account better characterizes the structure that comprehenders construct when reading sentences with RRCs.

2.2 The hypothesized structures for RRCs

A relative clause (RC) is a subordinate clause that modifies a noun. There are different types of relative clauses which differ in properties of the embedded clause.

For example, a passive reduced RC (RRC; such as (1), repeated below as (4)) has

the following properties: the modified noun is the subject of the embedded clause, the clause is in passive voice and does not contain an overt wh-phrase or finite auxiliary (i.e., it is *reduced*).

(4) The defendant examined by the lawyer was unreliable.

In a passive *full* RC (FRC; such as (2), repeated below as (5)), like in RRCs, the modified noun is the subject of the embedded clause and the clause is in passive voice. However, unlike a RRC, an FRC contains an overt wh-phrase and finite auxiliary.

(5) The defendant who was examined by the lawyer was unreliable.

On the other hand, a *progressive* reduced RC (ProgRRC; such as (3), repeated below as (6)) has all the properties of a RRC along with the following additional property: the embedded clause in a ProgRRC has a progressive aspect.

(6) The defendant being examined by the lawyer was unreliable.

We now describe two different accounts of the underlying structure of RRCs and how this structure relates to the structures of FRC and ProgRRC.

2.2.1 Two competing theoretical accounts of reduced RCs

The first hypothesis, which we will refer to as the *Whiz-Deletion hypothesis* (Chomsky et al., 1957; Ross, 1967; Smith, 1961), argues that both RRCs and FRCs encode all of the same information: when the structures of these two types of RCs are illustrated as trees, both the trees contain all of the same nodes (see Figure 2.1). The only difference between these two types of RCs is that the wh-phrase ("who") and the finite auxiliary

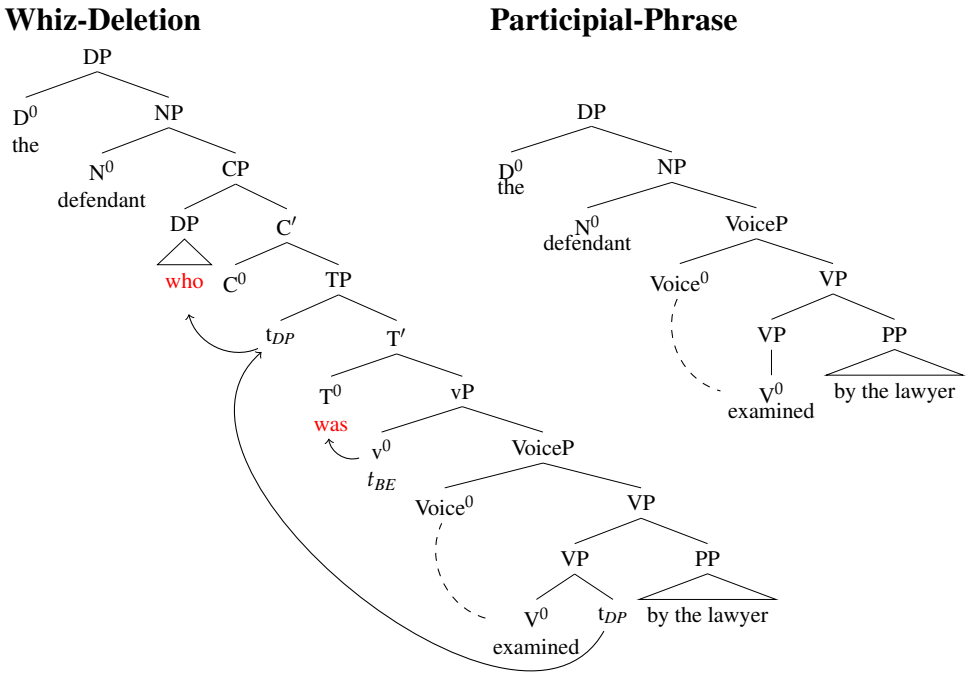


Table 2.1: Syntax trees to illustrate the structure of the **reduced** passive RC "the defendant examined by the lawyer..." under the Whiz-Deletion account (left) and Participial-Phrase account (right). The red words in the left tree are deleted and therefore not overtly produced. The structure of the **full** passive RC "the defendant who was examined by the lawyer ..." under both the accounts is illustrated by the left tree, but with the red words pronounced overtly. VoiceP is a short-hand for Voice Phrase. The assumptions involved in constructing these trees are described in the Appendix.

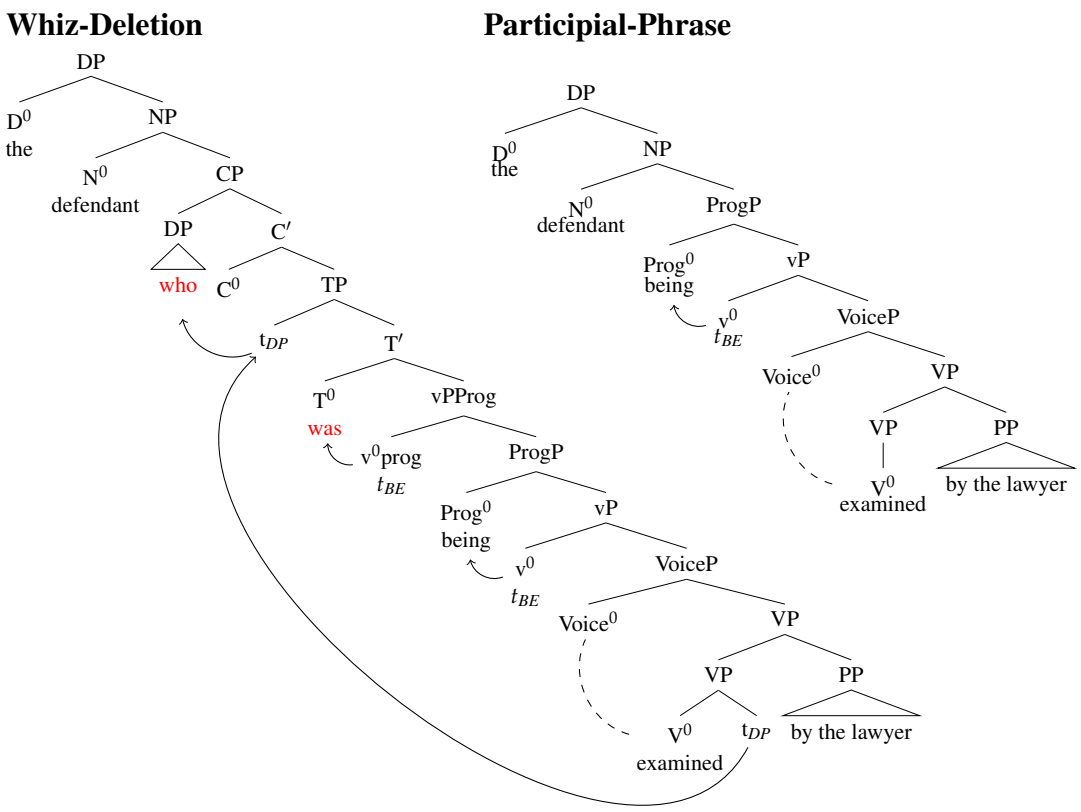


Table 2.2: Syntax trees to illustrate the structure of the **reduced** progressive RC "the defendant being examined by the lawyer..." under the Whiz-Deletion account (left) and Participial-Phrase account (right). The red words in the left tree are deleted and therefore are not overtly produced. The structure of the **full** progressive RC "the defendant who was being examined by the lawyer ..." under both the accounts is illustrated by the left tree, but with the red words pronounced overtly. VoiceP and ProgP are short-hands for Voice Phrase and Progressive Phrase respectively. The assumptions involved in constructing these trees are described in the Appendix.

("was") are deleted in RRCs at some point of syntactic processing, and are therefore not present in the surface structures. Under this account, ProgRRCs encode all of the information encoded in RRCs and FRCs as well as additional information that indicates that ProgRRCs have a progressive aspect: the nodes in the trees for FRCs and RRCs are a proper subset of the nodes in the tree for ProgRRC (see Figure 2.2). Some prior psycholinguistic work on RRCs has explicitly or implicitly assumed structures that are consistent with this account. For example, Tooley, Pickering, and Traxler (2019) explicitly state that full and reduced RCs "at an abstract level of representation ... have the same structure". Similarly, Fine and Jaeger (2016) implicitly assume that participants are constructing the same structure for FRCs and RRCs by assuming that repeated exposure to either of these types of sentences can cause RRCs to become surprising over the course of the experiment.

The second, more recent hypothesis, which we will refer to as the *Participial-Phrase hypothesis* (Harwood, 2018; Kayne, 1994; Bhatt, 1999), argues that FRCs contain a CP node which encodes that the clause introduces propositional content. This node is absent in RRCs and ProgRRCs. Like in the Whiz-deletion account, the nodes in the trees for RRCs are a proper subset of the nodes in the trees for ProgRRCs (compare Figures 2.1 and 2.2)

The motivation for the Participial-Phrase hypothesis comes from inflectional restrictions in English (such as the fact that relative clauses can be reduced in (7) but not (8)) compared to the existing Whiz-Deletion hypothesis.

(7) a. The defendant ~~who was~~ examined by the lawyer ...

b. The defendant ~~who was~~ examining the lawyer ...

c. The defendant ~~who was being~~ examined by the lawyer ...

(8) a. *The defendant ~~who will~~ examine the lawyer ...

b. *The defendant ~~who had~~ examined the lawyer ...

c. *The defendant ~~who might~~ have examined the lawyer ...

d. *The defendant ~~who will~~ be examining the lawyer ...

e. *The defendant ~~who had~~ been examining the lawyer ...

Based on evidence from other syntactic phenomena, Harwood describes principles which determine whether or not the CP node in an embedded clause can be omitted, the specific details of which are not relevant to this work. Then, he demonstrates that these principles can also explain the inflectional patterns described above, thus causing him to conclude that reduced passive and progressive relative clauses do not contain a CP node.

2.2.2 Generating hypotheses about representations used for incremental parsing under the two accounts

The trees in Figure 2.1 illustrate the information that comprehenders are expected to encode under the Whiz-Deletion and Participial-Phrase accounts, *after* they have parsed the sequence "the defendant examined by the lawyer". Since these accounts were not designed to explain incremental processing, it is not immediately transparent from these trees in Figure 2.1 what intermediate structures comprehenders are expected to construct under these two accounts *while* reading the sequence.

To articulate the hypotheses about the intermediate structures that comprehenders are expected to construct, we will use the Combinatorial Categorial Grammar (CCG; Steedman (1996)) formalism because, among other reasons we discuss in § 2.3.2.1, it is very straightforward to describe partially constructed structures using this formalism;

for work on incremental parsing with Minimalist grammars see Stabler (2013) and Baumann (2021) and with Context-Free grammars see Stolcke (1995) and Hale (2001).

We describe the CCG formalism in further detail in § 2.3.2.1, but the detail that is relevant for our current purposes is that in CCG the same lexical items (e.g., "defendant") can be associated with different syntactic categories (or *tags*) depending on the context the syntactic context they occur in. For example, "defendant" is associated with the tag *NP* when it is unmodified (as in (15)) but with the tag *NP/PP* when it is modified by a prepositional phrase (as in (10)); the "/" in the latter notation indicates that the noun *defendant* is part of a noun phrase (i.e., NP) that is waiting to be merged with a prepositional phrase (i.e., PP) on the right.

(9) The defendant examined the lawyer.

(10) The defendant with the binoculars examined the lawyer.

Taking advantage of these contextualized syntactic categories, we can describe the difference between the Whiz-Deletion and Participial-Phrase accounts in terms of the syntactic category that the nouns modified by RRCs, FRCs and ProgRRCs are associated with. Under the Whiz-Deletion account, in all of these types of RCs, the noun is associated with the tag *NP/CP* because the noun first combines with a CP: in the trees illustrated in Figures 2.1 and 2.2, the sister of the noun *defendant* (i.e., the node immediately to the right of the noun) is always a CP. In contrast, under the Participial-Phrase account, while the noun modified by a FRC first combines with a CP and therefore would be associated with *NP/CP*, the categories for nouns modified by RRCs and ProgRRCs are different: in RRCs, the noun first combines with a VoiceP, and would therefore be associated with *NP/VoiceP* (see Figure 2.1);

whereas, in ProgRRCs, the noun first combines with ProgP and would therefore be associated with *NP/ProgP* (see Figure 2.2).

Distinguishing between the accounts in this manner aligns with the incremental Minimalist parsing strategy used by Baumann (2021). In standard Minimalist Parsing, given a sentence like 'Grogru likes the movie', the parser first merges the verb *likes* with its direct object *the movie*. Then, this combined state merges with the subject of the sentence *Grogru*. This standard Merge operation is not compatible with an incremental parsing account because it does not allow the verb to merge with the subject before it merges with the object (i.e., the sequence *Grogru likes* cannot result in a merged state). Baumann introduces an incremental version of this merge operation ("forward merge"), which lets allows incomplete elements (such as a verb missing its direct object) to participate in merge operations. The "forward merge" operation is very similar to the forward-application rule in CCG which we introduce in § 2.3.2.1.

Why generate hypotheses from theories in generative syntax instead of just listing possible phrase structure hypotheses? The different NP types described above can also be expressed in terms of phrase structure rules as illustrated in Table 2.4. Given this, and the fact that theoretical accounts in generative syntax are not often designed to explain human sentence processing, the reader might wonder what is the advantage of generating hypotheses and predictions from these theoretical accounts instead of generating them by listing plausible phrase structure rules, as was done in some earlier psycholinguistic work studying structural representations (for a review of this work see Branigan and Pickering (2017)).

Coming up with possible phrase structure rules also requires starting with some

Whiz-Deletion	**Participial-Phrase**
NP → N	NP → N
NP → N CP	NP → N CP
NP → N PP	NP → N PP
NP → N and NP	NP → N and NP
NP → N PossessiveP	NP → N PossessiveP
	NP → N VoiceP
	NP → N ProgP

Table 2.3: Possible NP types under the Whiz-Deletion and Participial-Phrase account expressed as phrase structure rules.

theory about structure, otherwise it is not clear what the starting point for the phrase structure rules would be (cf. Gaston, Huang, and Phillips (2017)). If this theory is not grounded in a framework that is jointly trying to account for other phenomena (e.g., the structure of sentences across many if not all languages), then there are a lot of degrees of freedom about the level of abstraction at which to specify the phrase structure rules. For example, without any prior commitments about the structure of reduced RCs, any of the following rules can describe the noun phrase in RRCs: NP → N RC, NP → N RRC, NP→ N modifier, NP → animate_N RC, and so on. In many cases, it is not straightforward to select the correct rule amongst all of these possible rules merely based on empirical behavioral/ neural data for two reasons: first, existing behavioral/ neural methods are not sensitive enough to reliably pick up on very small differences (cf. Prasad and Linzen (2021)); second, even if existing (or new) methods were able to pick up on these small differences, the effects are often very sensitive to the specific experimental task being used (e.g., comprehension vs. production tasks) which makes it tricky to interpret any differences in effects across tasks without an explicit model of these experimental tasks (cf., Ziegler, Snedeker, and Wittenberg (2017)).

Given these concerns, we argue that when investigating what structures comprehenders construct when processing sentences, taking an independently motivated syntactic-theory first approach, combined with an explicit model of task demands (such as the one we introduce in the following section) can be a fruitful way to constrain the hypothesis space. For a review of prior psycholinguistic work that adopts such a syntactic-theory first approach and a discussion of when such an approach is fruitful, see Kush and Dillon (2021) and Phillips et al. (2021).

In the following section we specify a linking hypothesis between an incremental parser's parsing decisions and expected participant behaviour in a comprehension-to-production priming paradigm. Then, we describe the model of incremental parsing we developed and discuss how the different representational assumptions under the Whiz-Deletion and Participial-Phrase accounts are implemented in this model.

2.3 The Serial Parsing in ACT-R With Null-elements (SPAWN) model

The specific task we consider in this work is a web-based comprehension-to-production experiment (described in more detail in § 2.5). In the comprehension part of this task, participants read the prime sentences and re-type them from memory and in the production part participants complete ambiguous target prompts like "the participant examined". The manner in which participants complete the ambiguous prompt indicates how they parsed the prompt; if participants complete the sentence with a passive continuation (e.g., "the participant examined by the lawyer was unreliable") we can infer that participants parsed the noun in ambiguous parse as being in a RRC. We assume that there is no stochasticity in the production process: every time participants

assign a RRC parse to the ambiguous prompt, they will generate a passive continuation. Therefore, we assume that the probability with which participants produce passive completions in any priming condition is equal to the probability with which participants assign RRC parses to the prompt in that condition, as indicated in the equation below:

$$P(passive \mid prompt, prime) = P(RRC\ parse \mid prompt, prime) \tag{2.1}$$

To generate predictions for $P(RRC\ parse \mid prompt, prime)$ for different prime-target pairs under the Whiz-Deletion and Participial-Phrase accounts, we develop a serial parsing model whose parsing decisions are guided by principles of Adaptive Control of Thought - Rational (ACT-R; Anderson et al. (2004)). ACT-R is a cognitive architecture with integrated modules which are designed to explain general cognition through a small set of general computational principles and mechanisms that are relevant to a wide range of tasks and domains.

We do not fully implement our model using the ACT-R architecture because some of the modules, such as the perceptual motor module, are not directly relevant to the goals of this work. Instead, we use the relevant computational principles and mechanisms within the ACT-R framework that are relevant for parsing. Specifically, our model contains three components, following Reitter, Keller, and Moore (2011)'s ACT-R model of priming: declarative memory, which contains information about lexical and syntactic categories; procedural memory, which contains the algorithm for retrieving syntactic categories from memory and combining them together; and buffers which store the words the parser has encountered so far, the syntactic categories retrieved for those words and the current composed state of these retrieved categories.

In § 2.3.2 and § 2.3.3, we describe our implementation of declarative and procedural memory, including how our implementation accounts for the different representational assumptions made by the Whiz-Deletion and Participial-Phrase accounts. Before that, we discuss the difference between our proposed model and existing models of priming and parsing.

2.3.1 Prior models of parsing and priming

As described above, our model draws on some of the core implementational assumptions from Reitter, Keller, and Moore (2011)'s model of priming. However, there is one crucial difference between our model and that proposed by Reitter, Keller, and Moore: we model priming by modeling the process adopted by participants when *parsing* ambiguous sentences, whereas Reitter, Keller, and Moore model priming by modeling the process adopted by participants when *producing* sentences given a specified meaning, e.g., transfer of an object from an agent to a patient which can either be expressed using a prepositional object (X gave Y to Z) or a double object structure (X gave Z Y).

The experimental tasks that Reitter, Keller, and Moore model in which meaning needs to be pre-specified are not appropriate for studying phenomena like RRCs, where the two parses of an ambiguous prompt like "the defendant examined" have different meanings (i.e., *the defendant examined someone* and *the defendant was examined by someone*). Therefore, we propose a task (which we describe in § 2.5.1.3) where participants' productions are constrained not by a specified meaning, but rather by an ambiguous prompt. Our proposed model is better suited to model this experimental task. Additionally, since we model parsing, our model can generate predictions for

comprehension-to-comprehension priming paradigms, which Reitter, Keller, and Moore's model cannot do.

Our model also differs from the existing ACT-R model of parsing proposed by Lewis and Vasishth (2005) in three ways. First, SPAWN includes a mechanism for parsing structures with null elements, which is crucial for our research question since the core difference between the Whiz-Deletion and Participial-Phrase account lies in the null elements they assume (or do not assume). Second, for reasons outlined by Reitter, Keller, and Moore (2011) and in § 2.3.2, we implement structural knowledge in SPAWN using the CCG grammar (Steedman, 1996) formalism, whereas Lewis and Vasishth use a constituency based grammar. Third, the re-analysis mechanism used by the two models differ.[2] Lewis and Vasishth's model uses a repair-parsing in which previously discarded structures are re-activated based on a set of pre-specified repair operations (Lewis, 1993). Since we use a different grammar formalism, implementing these pre-specified repair operations in SPAWN is non-trivial. Therefore, as a starting step, in this chapter we implement re-analysis in SPAWN using an easier-to-implement backtracking strategy (described in § 2.3.3.2). For a helpful comparison of the two re-analysis strategies, see Lewis (1998).

There exist several symbolic and neural network based parsers that can incrementally parse temporarily ambiguous sentences (e.g., Hale 2001; Roark 2001; Ambati et al. 2015; Yang and Deng 2020). However, to the best of our knowledge, the parsing decisions of these parsers are not driven by specific cognitive principles such as the ones proposed by ACT-R. Consequently, the link between parsing and the mechanism underlying priming in any specific task in these models is not as transparent as in

[2] A re-analysis is necessary to parse temporarily ambiguous sentences in any serial-parser without an oracle which specifies the correct parsing decision at every step.

our proposed model. There also exist other models of priming that can generate predictions for priming in comprehension. However, these models either do not explicitly model syntactic structure (Chang, Dell, and Bock, 2006; Malhotra, 2009) or do not explicitly implement the priming mechanism Snider (2008) (for a more detailed discussion, see Reitter, Keller, and Moore (2011)).

We argue that in order to evaluate theoretical syntactic accounts which hypothesize that two sentences encode the same abstract properties, it is necessary to establish a transparent link between the proposed syntactic structures, the parsing algorithm and the priming mechanism. In the absence of such a link, it is not possible to interpret any potential null priming effects because there are two possible interpretations for any null effect: the absence of priming might be a consequence of the language processing/production system not being sensitive to the properties shared by the prime and target; or, the system might be sensitive to the shared properties, but these properties were not primeable under the experimental task (Ruiter and Ruiter, 2017; Koring and Reuland, 2017; Martin, Huetting, and Nieuwland, 2017; Rees and Bott, 2017; Ryskin and Brown-Schmidt, 2017). By specifying the mechanism for priming using ACT-R principles, we are making explicit assumptions about what properties of sentences can cause priming in a given task, thus making it easier to interpret null priming results.

2.3.2 Declarative memory

In the ACT-R framework, information in the declarative memory is stored in *chunks* which are bundles of attribute-value pairs. For example, a chunk can have an attribute called ‘color’ which can store the value ‘teal’. The type of a chunk is determined

by the specific attributes that it contains. In SPAWN we define two types of chunks: syntax chunks and lexical chunks. The attributes stored in these chunks are based on the Combinatorial Categorial Grammar (CCG) formalism Steedman (1996). In this section we first introduce the CCG formalism and motivate why we used it. Then, we describe the attributes present in syntax and lexical chunks. Finally, we highlight how the differences in the representational assumptions of the Whiz-Deletion and Participial-Phrase accounts are implemented as differences in the syntax chunks.

2.3.2.1 What is CCG and why do we use it?

CCG is a highly lexicalized grammar in which every word w is associated with a category which describes the syntactic categories of the words that w can combine with and the states that result from this combination. Complex categories are formed by combining simple categories with two types of functions: a left combining functor '/' and a right combining functor '\'. For example, transitive verbs are associated with the category (VP\DP)/DP, indicating that these verbs will first combine with a DP on the right (as indicated by '/DP') and then combine with a DP on the left (as indicated by '\DP') to result in a VP. There are a small set of combinatorial rules that determine the result of combining two categories together. One such rule is called the forward composition under which a category of the form X/Y combines with a category Y on its *right* to result in a category X. Another rule is called the backward composition rule under which a category of the form X\Y combines with the category Y on its *left* to result in a category X. These forward and backward composition rules drove the combination described in the transitive verb example above. There are only four other such rules that are licensed in the CCG formalism, which we list in Table 2.4 for completeness. For a justification of these rules, see Steedman (2001). It is not

necessary for the reader to understand this justification or know how to apply these in derivations in order to understand the remainder of this chapter.

Rule name	Parser state form	Tag form	Composed form
Forward composition	DP/NP	NP	DP
Backward composition	DP	TP\DP	TP
Forward harmonic composition	DP/VoiceP	VoiceP/PP	DP/PP
Backward harmonic composition	TP\DP	eos\TP	eos\DP
Forward crossed composition	CP/TP	TP\DP	CP\DP
Backward crossed composition	TP/VoiceP	eos\TP	eos/VoiceP

Table 2.4: Examples of all the six possible CCG composition rules being applied when parsing sentences in the training set.

We chose to use the formalism for two reasons. First, this formalism aligns closely with the Minimalist Program (Chomsky, 1995), which makes it straightforward to translate the representational assumptions from modern theoretical syntax into this formalism. For example, as discussed in § 2.2, the differences between the Whiz-Deletion and Participial-Phrase accounts can be described in terms of the syntactic category that is assigned to nouns that are modified by RRCs: NP/CP under the Whiz-Deletion account vs. NP/VoiceP under the Participial-Phrase account. This description closely aligns with complex lexical items in Minimalism that specify the arguments they are looking to merge with; for an incremental perspective on Minimalist derivations, see Baumann (2021).

The second reason for using CCG is that there exists a model of priming in production by Reitter, Keller, and Moore (2011) that also uses the CCG formalism. Implementing our model of parsing with a similar formalism makes it easier in the future to combine these two models and develop an integrated model of parsing and production. Generating and testing predictions from such a model can shed light

on whether it is feasible for the parsing and production systems to rely on the same underlying mechanisms (cf. Phillips (2013)).

2.3.2.2 Structure of the syntax and lexical chunks

Every syntax chunk in SPAWN contains four attributes: left, right, the combinator combining the two sides and the class of lexical items that can be associated with this chunk. The left and right keys can either contain simple or nested structures as illustrated below.

$$\begin{aligned} \{left &: DP, \\ right &: NP, \\ combinator &: /, \\ category &: det\} \end{aligned} \qquad \begin{aligned} \{left &: (TP\backslash DP), \\ right &: DP, \\ combinator &: /, \\ category &: verb\_trans\_act\} \end{aligned}$$

We will refer to syntax chunks in the remainder of this chapter using their category attribute; for example, we will refer to the chunk on the left as the *det* chunk, and the chunk on the right as the *verb_trans_act* chunk. For any syntax chunk, its corresponding CCG tag can be reconstructed by combining the left, right and combinator attributes; for example, the CCG tag for the *verb_trans_act* chunk would be (TP\DP)/DP.

Every lexical chunk contains two attributes, as illustrated below: the word form and the list of categories that constrains which syntax chunks this form can be associated with; a syntax chunk c can be retrieved when processing some word w only if label associated with the *category* attribute in c is present in the list associated with the

category attribute in w.

$$\{form : \text{`}the\text{'}, \quad category : [Det]\} \qquad \{form : \text{`}examined\text{'}, \quad category : [verb\_trans\_act, verb\_trans\_pass]\}$$

Note, the list of possible syntax chunks that can be associated with any word is hard-coded into the lexical representations with the *category* attribute in our current implementation for convenience; future work can explore alternative training paradigms in which the associations between lexical items and syntax chunks can be learned from corpus frequencies. Additionally, we do not include any additional semantic or syntactic features in our lexical chunks because these features are not important for the phenomenon we are studying. These features can be easily incorporated in future work as additional attribute-value pairs in the lexical chunk.

2.3.2.3 Differences between Whiz-Deletion and Participial-Phrase accounts

As discussed in § 2.2, the Whiz-Deletion account argues that all relative clauses (reduced or full) contain a CP node, whereas the Participial-Phrase account argues that reduced passive and progressive relative clauses do not contain this CP node. We implement this difference between the two accounts by creating two versions of declarative memory, one consistent with the assumptions of the Whiz-Deletion account and the other consistent with the assumptions of the Participial-Phrase account. In Table 2.5, we list the syntax chunks in both of these versions that are relevant for processing nouns and highlight which of these chunks are unique to either the Whiz-Deletion or Participial-Phrase versions of the declarative memory and which are

Functional category	**Syntax chunk**	**Whiz-Deletion?**	**Participial-Phrase?**
Unmodified nouns or nouns modified by adjectives	$\{left : NP$ $right :$ $combinator :$ $category : noun\}$	Yes	Yes
Nouns modified by a RC	$\{left : NP$ $right : CP$ $combinator : /$ $category : rc\_noun\}$	Yes	Yes (but not modified by RRCs or ProgRRCs)
Nouns modified by a RRC	$\{left : NP$ $right : Voice,$ $combinator : /$ $category : rrc\_noun\}$	No	Yes
Nouns modified by a ProgRRC	$\{left : NP$ $right : (VoiceP/ProgP)$ $combinator : /$ $category : progrrc\_noun\}$	No	Yes
Null wh phrase in subject gap	$\{left : CP$ $right : (TP\backslash DP)$ $combinator : /$ $category : null\_wh\_subj\}$	Yes	No
Null finite auxiliary	$\{left : (TP\backslash DP)$ $right : VoiceP$ $combinator : /$ $category : null\_finite\_aux\}$	Yes	No
Null progressive auxiliary	$\{left : (TP\backslash DP)$ $right : (VoiceP/ProgP)$ $combinator : /$ $category : null\_finite\_aux\}$	Yes	No
Null wh phrase in object gap	$\{left : CP$ $right : (((TP\backslash DP)/DP)/DP)$ $combinator : /$ $category : progrrc\_noun\}$	Yes	Yes

Table 2.5: Summary of the relevant syntax chunks for processing a noun under the Participial-Phrase and Whiz-Deletion accounts.

shared across both the versions.

In the Whiz-Deletion version, there are only two syntax chunks that nouns can be associated with: *noun* (for unmodified nouns or nouns modified by adjectives) and *rc_noun* (for nouns modified by relative clauses)[3]. In the Participial-Phrase version, nouns can be associated with two additional chunks: *rrc_noun* (for nouns modified by reduced passive RCs) and *progrrc_noun* (for nouns modified by reduced progressive RCs); unlike in the Whiz-Deletion version, in this version, reduced progressive and passive RCs are not associated with the *rc_noun* chunk. In order to make up for the missing *rrc_noun* and *progrrc_noun* chunks, the declarative memory in the Whiz-Deletion version consists of three null elements which are absent from the Participial-Phrase version: *null_wh_subj* (for the null wh-phrase in the subject position in RRCs and ProgRRCs)[4], *null_finite_aux* (for the null auxiliary in RRCs) and *null_finite_aux* (for the null auxiliarly in ProgRRCs).

The state instantiated by the *rrc_noun* chunk in the Participial-Phrase version (i.e. NP/VoiceP), can be derived in Whiz-Deletion version as follows: first, apply the forward application rule to combine the *rc_noun* chunk with the *null_wh_subj* to get the state NP/(TP\DP); then, apply forward-application rule again to combine this derived state with the *null_finite_aux* to get the state NP/VoiceP. Similarly, the state instantiated by the *rrc_noun* chunk (i.e., NP/ProgP) can be derived by combining

[3]We do not include chunks for other possible noun modifications in our current implementation, such as prepositional phrases, possessives or noun-noun compounds, because these chunks are not relevant for the sentences we will be considering to differentiate between the Whiz-Deletion and Participial-Phrase accounts. If these other types of modifications are relevant for future work, additional chunks can be easily added to the declarative memory.

[4]This chunk is different from *null_wh_obj*, which is the chunk associated with the null wh-phrase in the *object* position. As indicated in Table 2.5, *null_wh_obj* is present in both the Whiz-Deletion and Participial-Phrase versions, since Harwood (2018)'s argument against positing a covert CP structure applies only to RRCs and ProgRRCs, and not to reduced object RCs like "the defendant ~~that~~ the lawyer examined...".

rc_noun chunk with the *null_wh_subj*, and then combining the derived state with the *null_progressive_aux* chunk.

Deviation from standard grammar conventions in CCG Conventional analyses of relative clauses in the CCG formalism do not make use of the specific syntactic categories we have introduced such as NP/VoiceP or (TP\DP)/VoiceP. For example, in Hockenmaier and Steedman (2007)'s implementation of relative clauses, the noun that is being modified has the same category as unmodified noun (i.e., just NP). Instead, the words in the embedded clause combine together to form the category NP\NP (i.e., a category, which when combined with a NP on the right results in a NP), which is incidentally the same category as an adjective. Under this implementation, wh-phrases are essentially identity functions that take the category NP\NP and return NP\NP; since wh-phrases play no functional role, their absence in reduced RC does not change the derivation.

This implementation, while elegant, does not capture the representational assumptions made by either the Whiz-Deletion and Participial-Phrase account as described in § 2.2 and illustrated in Figures 2.1 and 2.2. For example, under both these accounts complementizers and wh-phrases are not just identity functions, but rather encode some functional information. For example, Adger (2003) describes the semantic role of complementizers as indicating "how the hearer should think of the proposition expressed by its clause: the main two possibilities are whether the clause should be thought of as a simple statement of fact, or a question about the facts". Under this definition, comprehenders might be expected to encode the lexical content in embedded clauses with and without CP nodes differently: only the content in clauses with a CP node would be encoded as separate proposition from the main clause. Therefore,

when the Whiz-Deletion and Participial-Phrase accounts differ on whether or not the structures of RRCs and ProgRRCs include a CP node, their disagreement is more substantive than a disagreement about the presence of null elements which do not play any functional role: they are disagreeing about how the content in these embedded clauses are encoded.

Since the goal of this work is to evaluate which of the two representational hypotheses about RRCs better describes the incremental structures than comprehenders build, it is necessary to faithfully model the differences between the representational assumptions. In order to have such a faithful implementation in our model, we define syntax chunks that deviate from standard CCG convention.

2.3.3 Procedural memory

In the ACT-R framework, the procedural memory consists of production rules which check if the contents of the buffers meet some condition, and if so, trigger actions that will add to or modify the contents of these buffers. In SPAWN, we specify three sets of production rules. The first set checks if the word currently being processed w_i has been assigned a syntax chunk and if not then these rules trigger from the declarative memory the retrieval of a syntax chunk c_{ij} drawn from the list of syntax chunks that can be associated with w_i. The second set checks if a chunk c_{ij} has been retrieved and if so triggers a process that combines the retrieved chunk with the structure that the parser has built after processing the previous $i-1$ words (i.e., the *parser state*; G_{i-1}). The final set checks if a combination was successful, and if not triggers a re-analysis of either the current or previous words.

In the remainder of the section we describe the retrieval mechanism used by the

first set of production rules and then describe the parsing algorithm that combines all the three sets of production rules.

2.3.3.1 Retrieval mechanism

When processing some word w_i, the parser retrieves from C_i (i.e., the set of possible syntax chunks associated with the w_i) the chunk which has the highest activation level. The activation level of a tag $c_{ij} \in C_i$ is based on the following formula specified by Anderson et al. (2004):

$$A_{ij} = B_{ij} + L_{ij} + S_{ij} + \epsilon \tag{2.2}$$

The first term of this equation, B_{ij}, is the base level activation of c_{ij}; this is similar to the prior probability of c_{ij} in a statistical parsing framework. The second term, L_{ij} is the lexical activation from the current word w_i to the chunk c_{ij}; this is similar to the conditional probability $P(c_{ij} \mid w_i)$ in a statistical parsing framework. The third term, S_{ij} is the spreading activation from the current parser state G_{i-1} to the chunk c_{ij}. The last term is random noise and is sampled from $Normal(0, \sigma)$; it allows for chunks with low activations to be occasionally retrieved. We describe below how B_{ij}, L_{ij} and S_{ij} are computed. The typical activation formula includes only three components: a base-level component, a context component and a noise component (for example, see Lewis and Vasishth 2005 and Reitter, Keller, and Moore 2011). We break down the context component into two parts by including L_{ij} and S_{ij}, because there are (at least) two ways in which the context can influence the activation.

Base-level activation B_{ij} is determined by two factors: first, the number of times c_{ij} has occurred in the sentences that the model has encountered;[5] second, the amount of time that has passed since the model encountered the sentences that contained c_{ij}. The base-level activation of a chunk is high if the model has encountered the chunk recently and/or frequently, as indicated in the formula below, where K indicates the total number of times the model has encountered c_{ij}, t_{ij_k} indicates the time since the model's k-th encounter of c_{ij}, and d is a decay parameter:

$$B_{ij} = \log \sum_{k=1}^{K} T_{ij_k}^{-d} \tag{2.3}$$

The term T_{ij_k} indicates the time taken to process all the words between the model's k-th encounter of c_{ij} and the current word w_i. This is specified by the equation below where i is the index of the current word, j_k is the index of the k-th word for which the model correctly retrieved c_{ij} for and t_l is the time taken to process some word w_l.

$$T_{ij_k} = \sum_{l=(i-j_k)}^{i} t_l \tag{2.4}$$

We compute t_l using the formula below from Vasishth and Engelmann (2021),

[5] Under this definition, the base-level activation of c_{ij} increases only when this chunk occurs in the final parse of the sentence, and not when the model retrieved this chunk but later discarded it during re-analysis. If we wanted to take all retrievals into consideration when computing the base-level activation, it is crucial to specify a cost term that penalizes the incorrect retrievals. In the absence of such a cost term, the model would not update its base-level activation to match the statistics of the environment as we might expect it to. To illustrate this point, let us consider an environment where c_ij occurs very frequently, and therefore a model trained in this environment has a very high base-level activation for c_{ij}. When the model is then put in a new environment where c_{ij} occurs infrequently, the rational behaviour is for the base-level activation of T to decrease over time in this environment. Since the base-level activation of c_{ij} in initially high, it will be often be retrieved incorrectly. The base-level activation will remain high if each of these misretrievals add to the original activation without some cost added. In order to avoid adding an additional cost hyper-parameter, we only consider correct retrievals when computing base-level activation.

where N is the number of chunks retrieved when processing w_l (this includes the chunks that the model retrieved every time w_l was re-analyzed), A_{ln} is the activation of the n-th chunk the model retrieved when processing w_l (as computed using Equation 2.2), F is a latency factor and f a latency exponent.

$$t_l = \sum_{n=1}^{N} F e^{-(f A_{ln})} \tag{2.5}$$

To illustrate how t_l is computed with a concrete scenario, consider the time taken to process the word *lawyer* in the sentence "the defendant examined the lawyer" when the parser makes the sequence of decisions listed in Table 2.6. Through the process of parsing the sentences, the parser considered five tags for the word *lawyer*. If we assume that the activations for these tags when they were retrieved were 1.5, 1, 1.1, 2.5 and 2, then the time taken to process *lawyer* is given by the formula below:

$$F \times (e^{-f \times 1.5} + e^{-f \times 1} + e^{-f \times 1.1} + e^{-f \times 2.5} + e^{-f \times 2})$$

Lexical activation The spreading activation from the word w_i to the chunk c_{ij} is proportional to the conditional probability of the chunk given the word, as indicated in the formula below where w is the current word being processed and M is the maximum activation that any item can spread.[6]

$$L_{ij} = M \times P(c_{ij} \mid w_i) \tag{2.6}$$

[6] An alternative approach would be to use the *fan* of the word, i.e., the number of items (e.g., in this case syntax chunks) that are associated with the word (cf. Lewis and Vasishth (2005)). Under this approach, all the items get equal weight, and this does not capture word specific biases like verb subcategorization biases. Therefore, we adopted the weighted spreading activation approach.

Stage	Word	Previously retrieved tags	Old goal state	Current tag	Goal state
1	The	[]	None	DP/NP	DP/NP
2	defendant	[]	DP/NP	NP/VoiceP	DP/VoiceP
3	examined	[]	DP/VoiceP	(TP\DP)/DP	(TP/VoiceP)/DP
4	the	[]	(TP/VoiceP)/DP	DP/NP	(TP/VoiceP)/NP
5	lawyer	[]	(TP/VoiceP)/NP	NP/CP	(TP/VoiceP)/CP
6	.	[]	(TP/VoiceP)/CP	end	FAILED
7	lawyer	[NP/CP]	(TP/VoiceP)/NP	NP/ProgP	(TP/VoiceP)/ProgP
8	.	[]	(TP/VoiceP)/ProgP	end	FAILED
9	lawyer	[NP/CP, NP/ProgP]	(TP/VoiceP)/NP	NP/VoiceP	(TP/VoiceP)/VoiceP
10	.	[]	(TP/VoiceP)/VoiceP	end	FAILED
11	lawyer	[NP/CP, NP/ProgP, NP/VoiceP]	(TP/VoiceP)/NP	NP	(TP/VoiceP)
12	.	[]	(TP/VoiceP)	end	FAILED
13	lawyer	[NP/CP, NP/ProgP, NP/VoiceP, NP]	(TP/VoiceP)/NP	NO TAGS	
14	the	[DP/NP]	(TP/VoiceP)/DP	NO TAGS	
15	examined	[(TP\DP)/DP]	DP/VoiceP	VoiceP/PP	DP/PP
16	the	[DP/NP, DP/NP]	DP/PP	DP/NP	FAILED
17	examined	[(TP\DP)/DP, VoiceP/PP]	DP/VoiceP	NO TAGS	
18	defendant	[NP/VoiceP]	DP/NP	NP	DP
19	examined	[(TP\DP)/DP, VoiceP/PP]	DP	(TP\DP)/DP	TP/DP
20	the	[DP/NP, DP/NP, DP/NP]	TP/DP	DP/NP	TP/NP
21	lawyer	[NP/CP, NP/ProgP, NP/VoiceP, NP]	TP/NP	NP	TP
22	.	[]	TP	end	SUCCESS

Table 2.6: One of the possible ways in which the parser implementing the Participial-Phrase account can end up with the correct parse for the sequence "The defendant examined the lawyer". The time taken to process each word depends on the number of tags that were retrieved in total during the parsing process. For example, there were five tags retrieved for the word "lawyer", at stages 5, 7, 9, 11 and 21. The activation level for each of these retrievals will influence the processing time for "lawyer" as indicated in Equation 2.5. FAILED indicates that the parser failed to combine the retrieved tag with the current parser state. NO TAGS indicates that the parser has tried to combine all the current parser state with all possible tags associated with the word and failed for all of them.

Spreading activation from the parser state The spreading activation from the current parse state G_{i-1} to the chunk c_{ij} is computed using the formula below where *Combine* is a function that returns true if c_{ij} can combine with G_{i-1}, V is the number of possible chunks that can be associated with w_i and can also be combined with G_{i-1}, $|.|$ denotes the size of a set and M refers to the maximum activation as in Equation 2.6.

$$S_{ij} = \begin{cases} \frac{M}{|V|}, & \text{if } \textit{Combine}(c_{ij}, G_{i-1}) \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

2.3.3.2 Parsing algorithm

As discussed earlier, our parser consists of three types of production rules. First, rules which retrieve a syntax chunk for some current word w_i using the mechanism described above. Second, rules which combine the retrieved chunk with the existing parser state by applying one of the six composition rules that are possible within the CCG formalism (listed in the Table 2.4); if either the parser state or retrieved chunk contains a nested rule (e.g., the rule for transitive verbs $(TP\backslash DP)/DP$), then this second set of production rules recursively apply the CCG composition rules to the nested parts (i.e., $(TP\backslash DP)$ in the transitive verb rule). Third, rules which trigger a re-analysis when the retrieved chunk cannot be combined with the current parser state. These rules retrieve a new chunk for w_i from the list of possible chunks that can be associated with w_i after excluding the originally retrieved chunk. If the list of possible chunks is empty, i.e., all of the possible chunks have already been retrieved and resulted in unsuccessful combinations, then the rules trigger re-processing of the previous word; this re-processing involves reverting the current parser state G_{i-1} to G_{i-2} (i.e., the state it was after processing w_{i-1}), and repeating the process of

Algorithm 1 Parsing algorithm

```
words ← sentence.split()
supertags ← [Null for word in words]
goals ← [Null]                          ▷ Holds parser goal states at each word
n ← len(words)
i ← 0
while i < n do
    combined ← Null
    word ← words[i]
    tags ← S[word]                      ▷ S[w] gives possible tags for w
    G ← goals[-1]
    j ← 0
    bad_options ← []                    ▷ Tags that can't combined with G

    while j < len(tags) do
        candidate ← generate_tag(word, G, tags, bad_options)
        combined ← combine(G, candidate)

        if combined not Null then
            supertags[i] ← candidate
            goals.add(combined)
            i ← i + 1                   ▷ Move to the next word
            break                       ▷ Stop looking for more tags
        else if combined is Null then
            bad_options.add(candidate)
            j ← j+1                     ▷ Move to the next tag
        end if
    end while

    if combined is Null then            ▷ Have not found a suitable tag
        goals.pop()                     ▷ Go back to previous parser state
        i = i-1                         ▷ Go back to previous word
        add_bad_tag(i, supertags[i])    ▷ Don't select same tag next round
    end if
end while
```

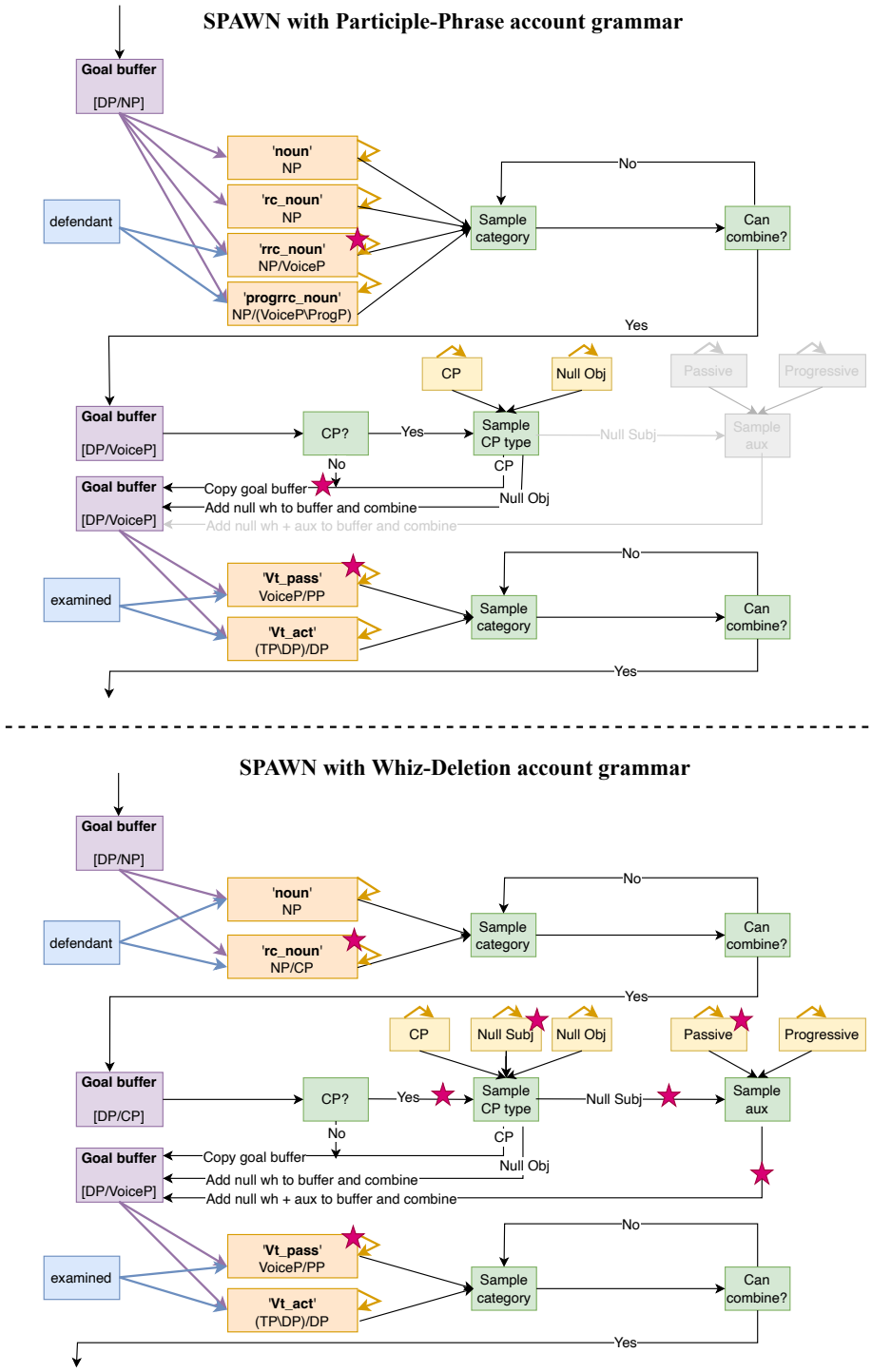


Figure 2.1: Visualization of the decisions that SPAWN has to make when processing the words "defendant" and "examined". Stars indicate the decisions that need to be made in order to select a RRC reading for this sequence. Greyed out portions indicate impossible paths. These paths are impossible because there is no chunk corresponding to a null wh-phrase in the subject position in the Participial-Phrase version of the declarative memory (see Table 2.5).

retrieving a syntax chunk w_{i-1}, but this time excluding c^*_{i-1} (the previously selected chunk for w_{i-1}) from the list of possible chunks. The parsing strategy involving these three sets of production rules is sketched in Algorithm 1.

Deviation from standard CCG parsing Standard approaches to CCG parsing consist of three steps (for an overview, see Clark 2021). First, retrieve categories for each lexical item (*supertagging*; Clark and Curran 2007; Xu, Auli, and Clark 2015; Tian, Song, and Xia 2020). Second, combine the categories using standard parsing algorithms (Hockenmaier and Steedman, 2002; Zhang and Clark, 2011). Third, find the parse with the highest score or probability. In this three step approach, the supertagging process can be independent of the combination process, which in turn can be independent of the process of selecting a single parse from the set of all possible parses. In contrast, in the parsing algorithm described above, the supertagging process is constrained by the combination process: when the parser is processing word w_i, only the categories or chunks that can combine with the current parser state G_{i-1} can be associated with w_i. Since the parsing algorithm is serial, at any given point i, only a specific sequence of possible chunks that can be associated with w_1 ... w_{i-1} is considered from the set of all possible sequences. The parser state G_{i-1} which constrains the chunk for w_i is a result of incrementally combining the chunk associated with the first word with the chunk associated with the second word, then combining the result with the chunk associated with the third word and so on.

This incremental combination presents a problem for CCG parsing because sometimes, depending on the grammar we adopt, it is necessary for words w_{i+1} and w_{i+2} to combine with each other before they can combine with w_i. To illustrate this, let us consider a simple grammar with just four words:

- *Grogru*: DP
- *likes*: (TP$\backslash$DP)$/$DP
- *the*: DP$/$NP
- *movie*: NP

When parsing the sentence "Grogru likes the movie" using this grammar in a bottom-up manner, the following steps are involved:

1. Combine "the" and "movie" by applying forward composition.
   "the movie": DP
2. Combine "likes" and "the movie" by applying forward composition.
   "likes the movie": (TP$\backslash$DP)
3. Combine "Grogru" with "likes the movie" by applying backward composition.
   "Grogru likes the movie": TP

Therefore, in this example, "likes" first combines with "the movie" before it combines with "Grogru". However, since our parsing algorithm combines words as it encounters them, "likes" needs to be able to first combine with "Grogru". A common approach to make such incremental combination possible is to introduce *type-raising* rules which change the syntactic category of words into new categories that are looking to combine with the original category. For example, consider the following type-raising rule which changes the category of words associated with DP, such as "Grogru" or "the movie", into a category in which a TP is looking to combine with a DP: DP $\Rightarrow_{\mathbf{T}}$ TP/(TP$\backslash$DP). By type-raising "Grogru" to TP/(TP$\backslash$DP),

it becomes possible to combine "Grogru" with "likes" using the forward composition rule to result in the following combined state for "Grogru likes": TP/DP.

As a starting point, we approximately implemented type-raising in our model by recursively applying CCG composition rules to nested parts of chunks and/or parser states. For example, let us again consider the scenario where the parser tries to incrementally combine "Grogru" with "likes". Since the chunk associated with "Grogru" (i.e., DP) cannot combine with the chunk associated with "likes" (i.e., (TP\DP)/DP), recursive rule application is triggered: first, the parser applies the backward composition rule to combine DP with the nested part (TP\DP), resulting in the nested combined state TP; next, the parser replaces the nested part in the original rule with the nested combined state, resulting the following combined state for "Grogru likes": TP/DP. Thus, the recursive combination results in the same combined state as applying type-raising in most cases.

Approximately implementing type-raising in this manner makes it slightly easier to specify the procedural and declarative memory components: we do not need to add additional production rules to the procedural memory which specify when a retrieved chunk can be type-raised, or additional syntactic chunks in the declarative memory for already type-raised chunks. The disadvantage is that this approximate implementation of type-raising is not restrictive enough, and sometimes results in application of type-raising rules that are not valid in CCG because they are not "order preserving" (Steedman, 1991). The application of these invalid type-raising rules causes the parser to sometimes construct intermediate parser states that can never be successful. The practical consequence of this is that additional re-analyses get triggered when the parser constructs these invalid states, thus resulting in an overestimate of the time

taken to process and word or sentence. Since our goal was not to model the time taken to process sentences, but the final syntax chunks associated with words in the sentence, we did not expect the construction of these invalid states to qualitatively affect our results. However, for more precise predictions, it is necessary to implement an adequately restrictive version of type-raising in future work.

2.3.3.3 Strategy for dealing with null elements

For any given parser state G_{i-1}, the parser checks if the most recently retrieved chunk was *rc_noun* (i.e., NP/CP). If so, then the production rules trigger a sampling process where the parser decides whether this chunk will be followed by an overt complementizer (like in full RCs), by a null object complementizer (like in object reduced RCs such as (11)) or by a null subject complementizer (like in RRCs and ProgRRCs under the Whiz-Deletion accont).

(11) The defendant ~~who~~ the lawyer examined was unreliable.

If the parser samples the overt complementizer, i.e., the parser expects to encounter a complementizer without any bottom-up cues, then the parser moves on to process the next word without taking any actions. On the other hand, if the parser samples one of the null complementizers, then the parser retrieves this complementizer and combines it with the current parser state. In cases where the parser samples a null *subject* complementizer, the parser additionally samples one of two types of auxiliaries — passive and progressive — and then combines these with the parser state before processing the next word. The probability with which a complementizer or auxiliary chunk is retrieved is proportional to the base level activation of the chunk (as defined in Equation 2.3). Note, since a chunk for the null subject complementizer only exists

	Word	Previously retrieved tags	Old goal state	Current tag	Goal state
1	The	[]	None	DP/NP	DP/NP
2	defendant	[]	DP/NP	NP/CP	DP/CP
3	null-wh	[]	DP/CP	CP/(TP\DP)	DP/(TP\DP)
4	null-aux	[]	DP/(TP\DP)	(TP\DP)/VoiceP	DP/VoiceP
5	the	[]	DP/VoiceP	DP/NP	FAILED
6	the	[DP/NP]	DP/VoiceP	NO TAGS	
7	null-aux	[(TP\DP)/VoiceP]	DP/(TP\DP)	(TP\DP)/(VoiceP/ProgP)	DP/(VoiceP/ProgP)
8	the	[DP/NP]	DP/(VoiceP/ProgP)	DP/NP	FAILED
9	the	[DP/NP, DP/NP]	DP/(VoiceP/ProgP)	NO TAGS	
10	null-aux	[(TP\DP)/VoiceP, (TP\DP)/(VoiceP/ProgP)]	DP/(TP\DP)	NO TAGS	
11	null-wh	[CP/(TP\DP)]	DP/CP	CP/(((TP\DP)/DP)/DP)	DP/(((TP\DP)/DP)/DP)
12	the	[DP/NP, DP/NP DP/NP]	DP/(((TP\DP)/DP)/DP)	DP/NP	DP/(((TP\DP)/DP)/NP)
13	lawyer	[]	DP/(((TP\DP)/DP)/NP)	NP	DP/(((TP\DP)/DP))
14	examined	[]	DP/(((TP\DP)/DP))	((TP\DP)/DP)	DP
15	was	[]	DP	((TP\DP)/(NP\NP)	TP/(NP\NP)
16	unreliable	[]	TP/(NP\NP)	NP\NP	TP
17	.	[]	TP	end	SUCCESS

Table 2.7: One of the possible ways in which the parser implementing the Whiz-Deletion account can end up with a correct parse for the sentence "The defendant the lawyer examined was unreliable". FAILED indicates that the parser failed to combine the retrieved tag with the current parser state. NO TAGS indicates that the parser has tried to combine all the current parser state with all possible tags associated with the word and failed for all of them.

in the Whiz-Deletion version of SPAWN, this chunk can never be retrieved by the Participial-Phrase version. This difference between the two accounts is illustrated in Figure 2.1. Note, any differences between the Participial-Phrase and Whiz-Deletion versions of SPAWN are driven by the differences in the set of assumed syntax chunks (see Table 2.5); the production rules are identical in both the versions.

Re-analysis with the null elements For the purposes of re-analysis, null elements are treated like other syntactic categories: re-analysis of the previous words is triggered only after all relevant null-element tags are considered and discarded. For example, consider a scenario where the parser is processing a reduced object RC and the parser

selects *null_wh_subj* and then the finite auxiliary, thus expecting the upcoming word to be a verb in a reduced RC (rows 3 and 4 in Table 2.7). Instead, the parser encounters a determiner (row 5), indicating that the parser's decisions to select the *null_wh_subj* and/or the finite auxiliary were incorrect. At this stage, the parser goes back one step in the decision process and selects the progressive auxiliary, now expecting the upcoming word to be "being" (row 7). Since this expectation is inconsistent with the determiner it encounters (row 8), the parser now goes back two steps and selects *null_wh_obj* (row 11), and is then able to successfully process the determiner.

Algorithm 2 Null element algorithm

```
if CanHaveNull(state) then        ▷ Check if current state can generate null element
    next_cat = SampleNext(state)                          ▷ Retrieve the next category

    if IsNullEl(next_cat) then                  ▷ Check if next_cat is a null element
        combined = combine(state, next_cat)
        state = combined
        continue                                                  ▷ Move to next word
    else
        continue              ▷ If next_cat is not a null element move to next word
    end if

else
    continue                  ▷ If state can't generate null element move to next word.
end if
```

How generalizable are the proposed production rules for sampling null-elements? The production rules we proposed for sampling null-elements are specific to relative clause processing in the sense that the specific null-elements being sampled — null wh-phrases and null-auxiliaries — are not relevant when processing other sentences. However, the general strategy we proposed is applicable to all contexts in which

null-elements can occur. For example, consider the following examples in which the red and striken-through words can either be deleted or not.

(12) The defendant examined the evidence and the **lawyer** ~~examined~~ the defendant. (Gapping)

(13) The defendant can examine the evidence and the lawyer **can** ~~examine the evidence~~ too. (Verb-Phrase Ellipsis)

(14) The defendant examined something but the lawyer couldn't see **what** ~~the defendant examined~~. (Sluicing)

In all of these examples, after an incremental parser processes the bolded word in the sentence, it needs to predict whether or not the words will be elided in the sentence. If the parser predicts that the words will be elided, then the parser needs to trigger a retrieval mechanism that will retrieve the words that it predicts will be elided, and then combine them with the current goal state. On the other hand, it the parser predicts that the words will not be elided, then it can continue processing the words without any further action. We illustrate this general strategy for parsing sentences with null elements in the Algorithm 2.

2.3.4 What factors cause priming in the SPAWN models?

As discussed earlier, we assumed that there was no stochasticity in participants' productions — i.e., given a target prompt which is ambiguous between a main verb and reduced RC readings (e.g., "the defendant examined"), the probability with which participants produce continuations consistent with the reduced RC parse is equal

to the probability with which participants assign a reduced RC parse to the prompt (Equation 2.1). Given this assumption, the factors that can cause priming in our modeling setup are the factors that can influence the probability with which the Whiz-Deletion and Participial-Phrase versions of the SPAWN models select a reduced RC parse under different priming conditions.

In our implementation of SPAWN, the parse that the model assigns to the ambiguous target prompt is determined by the syntax chunk that the model associates with the noun in the prompt (e.g., *defendant* in the prompt "the defendant examined"). In the Participial-Phrase version, a RRC parse is assigned if the model retrieves the *rrc_noun* chunk. In the Whiz-Deletion version, on the other hand, a RRC parse is assigned only when the model first retrieves the *rc_noun* chunk and then subsequently retrieves the *null_wh_subj* and *null_finite_aux* chunks (see Figure 2.1).

The probability of some chunk c_{ij} being retrieved when the parser is processing some word w_i is influenced by the activation of c_{ij} which in turn is influenced by four factors (as specified in Equation 2.2): the base level activation of c_{ij}, the lexical activation from w_i to c_{ij}, the activation from the current parser state G_{i-1} to c_{ij}, and noise. In the prime-target pairs we consider (described in detail in § 2.5.1.2), the activation from the target noun w_i to the relevant noun chunks c_{ij} does not differ across the priming conditions because there is no noun overlap between the prime and target pairs; since the target noun w_i does not occur in the primes, processing the prime sentences will not change the activation w_i spreads to c_{ij}. Similarly, the activation from G_{i-1} to c_{ij} also does not differ across priming conditions because there is only one possible state that the parser can be in when processing the target noun, which is DP/NP; therefore, processing the primes will not change G_{i-1} and

hence will not change the activation it spreads to c_{ij}.

Given these properties of the prime-target pairs we consider, any differences we observe across the priming conditions has to be driven by the differences in the base-level activation of the relevant noun chunks. We outline below how processing the prime sentences in different conditions can influence these base-level activations.

Priming in the Whiz-Deletion version As discussed above, a RRC parse is selected by the Whiz-Deletion version when the model first retrieves the *rc_noun* chunk when processing the target noun followed by the retrieval of the *null_wh_subj* and *null_finite_aux* chunks. Crucially, given our implementation of re-analysis described in § (11), when the model is parsing an ambiguous target prompt, as long as the model retrieves the *rc_noun* chunk, it is guaranteed to also eventually retrieve the *null_wh_subj* and *null_finite_aux* chunks. Therefore, any prime sentence that increases the base-level activation of the *rc_noun* chunk will increase the probability of the model assigning a RRC parse for the subsequent target. Since the *rc_noun* chunk occurs in all RC sentences under the Whiz-Deletion account, we would expect RRC, ProgRRC and FRC primes to increase the base-level activation of this chunk to a larger extent than a minimally different active sentence without a relative clause (like (15)).

(15) The defendant examined the lawyer and was unreliable. (Active Main Verb; AMV)

Therefore, the following equation describes the expected probability of passive target continuations that are consistent with a RRC parse (i.e., $P(pass)$) under the different priming conditions with the Whiz-Deletion version of SPAWN.

$$P(pass \mid RRC) = P(pass \mid ProgRRC) = P(pass \mid FRC) > P(pass \mid AMV) \tag{2.8}$$

Priming in the Participial-Phrase version As discussed above, a RRC parse is selected by the Participial-Phrase version when the model retrieves the *rrc_noun* chunk when processing the target noun. Since the *rrc_noun* chunk occurs only in RRC sentences, we would expect only RRC primes, but not ProgRRC or FRC primes, to increase the base-level activation of this chunk relative to minimally different AMV primes. Therefore, the following equation describes the expected probability of passive target continuations that are consistent with a RRC parse (i.e., $P(pass)$) under the different priming conditions with the Participial-Phrase version of SPAWN.

$$P(pass \mid RRC) > P(pass \mid ProgRRC) = P(pass \mid FRC) = P(pass \mid AMV) \tag{2.9}$$

The need for generating quantitative predictions Equations 2.8 and 2.9 describe the qualitative patterns of results we expect to find by reasoning about the computational principles of the Participial-Phrase and Whiz-Deletion versions of the SPAWN model. In the next section, we generate quantitative priming predictions by presenting our models with the specific prime sentences and target prompts in the specific order they were presented in the empirical experiment described in § 2.5. There are two advantages to generating these quantitative predictions.

First, the computational processes of the SPAWN models, while interpretable,

can interact in complex ways. Therefore it is possible that we failed to consider the consequences of some of these interactions when generating the predictions by reasoning about these principles. Generating quantitative predictions can thus serve as a method of validating the qualitative predictions.

Second, the specific hyperparmeter setting we used for the SPAWN models (which we described in § 2.4.1.2 and will describe in further detail in the following section), while reasonable, is not the only possible setting we could have used. By generating quantitative predictions, we can evaluate these decisions by comparing the predictions to the magnitude of effects observed in the empirical human data.

2.4 Generating quantitative priming predictions under the different accounts

2.4.1 Methods

2.4.1.1 Training data

In order to approximately model the base-level and lexical activations participants might come in to the experiment with, we templatically generated a dataset with 10000 sentences. Only about 6.5% of the sentences in this dataset contained relative clauses. Table 2.8 lists the structures included in this dataset, the probability with which these structures occurred, and an example sentence for each structure.

To estimate the probabilities for each of the RC structures in our dataset, we used corpus frequencies estimated by Roland, Dick, and Elman (2007). The frequencies of sueject RC, full and reduced object RCs and passive RCs were directly estimated by Roland, Dick, and Elman using both spoken and written corpora. For these structures, we estimated the probabilities by averaging across the frequencies in the three written

Structure	Prob	Example
Subject RC	0.016	The defendant who examined the lawyer ...
Full object RC	0.002	The defendant who the lawyer examined ...
Reduced object RC	0.005	The defendant the lawyer examined ...
Full passive RC	0.002	The defendant who was examined by the lawyer ...
Reduced passive RC	0.011	The defendant examined by the lawyer ...
Full progressive RC	0.0002	The defendant who was being examined by the lawyer ...
Reduced progressive RC	0.0002	The defendant being examined by the lawyer ...
Transitive NP object	0.321	The examined the lawyer.
Transitive PP object	0.080	The defendant went to the store.
Intransitive	0.241	The defendant sang (joyfully).
Copular	0.241	The defendant was happy.
Coordination	0.080	The defendant examined the lawyer and went to the store. The defendant was happy and sang joyfully. The defendant went to the store and sang and was happy and examined the lawyer.

Table 2.8: Structures that were present in the templatically generated training dataset.

corpora: the Wall Street Journal, the Brown Corpus and the British National Corpus. Roland, Dick, and Elman did not estimate the probability of progressive passive RCs. Therefore, we used the frequencies of passive infinitive relative clauses such as "The last defendant to be examined in the court ..." to estimate the probability for progressive passive RCs and divided this probability equally between the reduced and full versions of progressive passive RCs. We used passive infinitive relative clauses because they shared two properties with progressive passive RCs: first, the embedded clause was in passive voice; and second, the embedded clause contains a 'be' auxiliary.

For the sentences without RCs, we wanted to include some variety in the sentences structures, but did not expect the exact probabilities of these sentences to matter for our purposes. So we divided the remaining probability mass (93.5%) based on the approximate expected frequency of these structures.

To generate sentences with RCs, we first generated the RC by sampling a subject (which was always animate), verb and object for the embedded clause, and then

sampled one of the five structures without RCs to generate the main clause. For sentences with coordination we sampled two sentences without RCs based on their relative probabilities. Since there was a non-zero chance of coordination being sampled again, this allowed for nested coordinated sentences. Further details about the template can be found on Github.[7]

The lexical items that filled the slots in the templates were sampled from a set of 404 lexical items: 215 nouns (163 animate), 110 verbs (39 that could result in MV/RR ambiguity and occurred in embedded clauses; 5 took adjectives as complements and occurred in the copular condition), 17 determiners or bare nouns that could occur as DPs, 26 adjectives, 29 adverbs and 7 prepositions. Some of the experimental sentences contained some structures that could not be parsed with our CCG grammar. For example, our grammar did not contain the syntactic categories required to parse complex adverbs like "ran away in fear" in the sentence "The thief identified by the victim ran away in fear". To over come this, we created multi-word lexical items like an intransitive verb "ran-away" and an adverb "in-fear".

As discussed in § 2.3.4, since nouns were not repeated across the primes and targets in our experimental design, priming of SPAWN models will be driven only by base-level activations and not by factors like lexical frequency or semantic plausibility. Therefore in the current implementation, we did not take these factors into consideration when sampling lexical items.

2.4.1.2 Hyperparameters

There are two types of hyperparameters in the SPAWN models, ones which differ across participants and ones which don't. The following three hyperparmeters are

[7] https://github.com/grushaprasad/spawn

fixed across all instances of the model, and were set based on prior work (Lewis and Vasishth, 2005)

- Decay parameter (d in Equation 2.3): the speed at which activation decays; set to 0.5.
- Latency exponent (f in Equation 2.5): the rate at which activation of a tag influences the time taken to retrieve the tag; set to 1.
- Maximum activation (M in Equations 2.6 and 2.7): the amount of activation that can spread from any chunk; set to 1.5.

The following three hyperparameters differ across differenct model instances.

- Noise parameter (σ): the standard deviation of the distribution from which ϵ (see 2.2) is sampled; a different value of this parameter is sampled for each model instance from $Uniform(0.2, 0.5)$ as per Vasishth and Engelmann (2021).
- Latency factor (F in Equation 2.5): the rate at which the exponent of activation of a tag influences the time taken to retrieve the tag; a different value sampled for each model instance from $Beta(2, 6)$ as per Vasishth and Engelmann (2021).
- Random seed: this influences the sentences each model instance is trained on, the order of these training sentences, as well as the stochasticity in the parsing process.

2.4.1.3 Procedure

We trained 1280 different instances of both the Participial-Phrase and Whiz-Deletion versions of the SPAWN model on 100 sentences from the training dataset by varying the hyperparameters described in the previous section. Such a high number was required to ensure that we could find unambiguous evidence for the differences or equivalence in the probability of RRC parses across conditions as estimated using Bayes Factors (Jeffreys, 1939). We describe this further in § 2.4.2.1. For any given instance, the hyperparameters were fixed across the Participial-Phrase and Whiz-Deletion version. During training the models parsed each training sentence and updated the base-level and lexical activations based on the tags assigned to the individual words in the sentence.

The fact that the models were trained on 100 sentences reflects our assumption that participants will only only weakly weight their prior experience when parsing sentences in the experiment. While it is implausible that only 100 sentences influence participants' beliefs about the distribution of sentences, prior work has also used such small numbers (e.g., 178 by Fine et al. 2010 and 77 by Delaney-Busch et al. 2019) because they were effective in empirically estimating the strength of participants' prior beliefs. When we trained models on 500 sentences, only 2 out of the 1280 model instances ever assigned a RRC parse to any of the ambiguous target prompts, suggesting that it is necessary for us to assume weak prior knowledge.[8]

We assigned each of these model instances to one of 32 experimental lists which contained 24 items (6 per priming condition). Each item consisted of three prime

[8]Even when we added more noise to the models' estimates by sampling σ from $Normal(0.35, 1)$ instead of from *Uniform(0.2,0.5)*, only 84 out of 30720 target prompts across all model instances were assigned a reduced RC parse.

sentences followed by one ambiguous target prompt. The target verb was repeated across the items but the subject noun was not. The lists are described in further detail in § 2.5.1. In the experiment simulation, for every item, the model parsed the three prime sentences and updated the activations, like during training. Then, it generated a sequence of tags for the partial prompt. We constrained the parser so that the parser could only end of one of two states that could eventually result in a grammatical sentence: TP/DP (consistent with a MV reading) or DP/PP (consistent with a RRC reading).[9]We then recorded, for every item, whether the model assigned a passive or active tag to the verb.

2.4.2 Predictions

2.4.2.1 Statistical model

To generate quantitative predictions about the predicted proportion of passive responses while taking into consideration the model-instance wise and item wise variation, we fit two Bayesian mixed effects logistic regression models with different contrasts. In logistic regression, the probability of some event *e* having one of two outcomes — e.g., in our case the probability of the target prompt receiving either a reduced RC reading or a main-verb reading — is modeled by expressing the *log-odds ratio* of that event as a linear combination of predictions. The log-odds ratio of the target prompt receiving a reduced RC parse in some condition *C* is given by the following formula, where *p* refers to *P(RRC parse | target, C)*:

[9]Such a constraint was not required when the parser was parsing full sentences because the later words in the sentence allowed the parser to discard all the incorrect incremental states it built along the way.

$$\textit{Log-odds(RRC parse} \mid \textit{target, C)} = \log\left(\frac{p}{1-p}\right)$$

We fit two models with two sets of contrasts to evaluate all of the directional predictions described in Equations 2.9 and 2.8. In the first model, we used a baseline coding scheme where we compared the mean log odds ratio of the AMV condition to the mean log odds ratio in each of the other conditions; this let us evaluate if the mean log odds ratio of AMV is greater than the mean log odds ratio of the ProgRRC and FRC conditions (as in Equation 2.8) or equal (as in Equation 2.9). In the second model, we used a Helmert contrast coding scheme with the following predictors, which let us evaluate if the mean log odds ratio of the ProgRRC and FRC conditions are equal to each other and to the mean log odds ratio of the RRC condition (as in Equation 2.9).

- C1: Compare the mean log odds ratio of the AMV condition with the mean log odds ratio of all the RC conditions combined.
- C2: Compare the mean log odds ratio of the RRC condition with the mean log odds ratio of all ProgRRC and FRC conditions combined.
- C3: Compare the mean log odds ratio of the ProgRRC condition to the mean log odds ratio of the FRC condition.

For both of these contrasts, we fit the maximal model by including all by-model-instance and by-item random intercepts and slopes, as described below. We use the same models to analyze the data from the human experiments described in § 2.5.

Model 1:

$$Passive \sim rrc + progrrc + frc +$$
$$(1 + rrc + progrrc + frc \mid item) +$$
$$(1 + rrc + progrrc + frc \mid instance)$$

Model 2:

$$Passive \sim c1 + c2 + c3 +$$
$$(1 + c1 + c2 + c3 \mid item) +$$
$$(1 + c1 + c2 + c3 \mid instance)$$

Finding evidence for either of the two directional predictions involves finding evidence for a null effect: for the Whiz-Deletion version, this involves finding that the difference between ProgRRC and FRC conditions (as measured using C2 in the Helmert coded model) is 0; for the Participial-Phrase version, this involves finding that the differences between AMV and both ProgRRC and FRC (as measured in Model 1) are 0. We will conclude that there is evidence for null effects if the Bayes Factor is less than $1/3$ (Jeffreys, 1939). Since Bayes Factors are very sensitive to the prior distribution (Schad et al., 2022), where uninformative priors bias the model towards a null effect, we fit the model using the following weakly informative prior.

$$\textit{Intercept} \sim Normal(-4.595, 1.5)$$

$$\textit{Fixed effects} \sim Normal(0, 2)$$

$$\textit{SD for random effects} \sim Normal(0, 5)$$

This prior assumes that the log odds ratio between priming conditions is most likely to be 0 (i.e. no priming effect) and unlikely to be greater than 4 or less than -4. This assumption is based on a meta-analysis of priming in production studies (Mahowald et al., 2017) where the log odds ratio between the prime conditions was not greater than 4 in any of the constructions they considered.

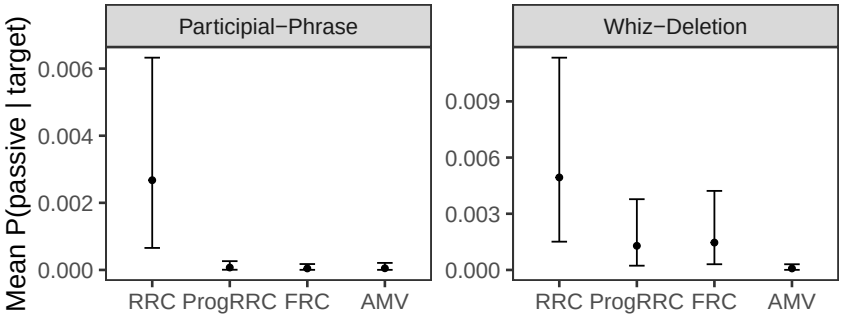
In the resulting prior predictive distribution, the predicted proportion of reduced RC sentences in the experiment ranges from 0 to 0.56 and the 95% quantile is 0.02—0.44. The log odds for each of the three specified contrasts in the prior distribution is centered at 0 and mostly falls between -5 and 5; this means that under this prior, effects in the predicted direction are as plausible as effects in opposite direction of what we might expect (e.g., a greater proportion of passive responses in the AMV condition compared to the RRC condition). Thus, the priors we specified are not overly restrictive, which is a desirable property in general, but especially so in our case because we are analyzing data from a novel model and a novel experimental paradigm.

2.4.2.2 Results

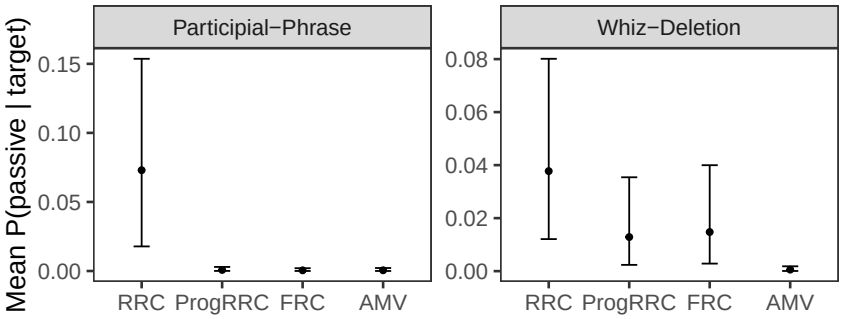
We list the model estimates and Bayes Factors in Table 2.9. In Figure 2.2, we plot the probability of the models assigning a passive tag to the target verb (i.e., selecting a RRC parse) under the different priming conditions as predicted by the Participial-Phrase and Whiz-Deletion accounts. To compute these probabilities, we sample the mean log odds ration for each of the contrasts from the posterior distribution of the Helmert coded model specified in 2.4.2.1. From these mean log odds values of the contrasts, we computed the mean log odds ratio for each of the priming conditions, and then transformed this log odds ratio into probabilities.

Even with 1280 model instances, the Bayes Factors indicated that there wasn't sufficient evidence in support of either the null or alternative hypothesis for some comparisons (see Table 2.9). This was a consequence of extremely low predicted probabilities, which in turn was a result of the fact that most model instances never generated a RRC parse. Since there cannot be any effect of priming in model instances that never generated a RRC parse, we repeated the analyses by including only the model instances which assigned a passive parse to at least one of the target prompts (Passive-Models). The qualitative pattern of results was identical across the analysis with all model instances and the analysis with only Passive-Models (see Figure Figure 2.2). However, in the Passive-Models analysis, there was moderate to strong evidence in support of either the null or the alternative hypotheses for all of the comparisons of interest (see Table 2.9).

Under both accounts the mean probability of the target prompt being assigned a RRC parse was highest in the RRC priming condition. However, even in this condition, the probability values were extremely low: the mean P(RRC parse $\mid$ target) in the RRC



(a) Predicted probabilities with all model instances (N=1280)



(b) Predicted probabilities (Participial-Phrase N=377; Whiz-Deletion N=450)

Figure 2.2: Probability of the target verb being assigned a passive tag (i.e., selecting a RRC parse) given an ambiguous target prompt with the Participial-Phrase and Whiz-Deletion versions of the SPAWN model. Plots were generated from the posterior samples of the Helmert coded Bayesian mixed effects model (described in 2.4.2.1). Error bars reflect 95% Credible Intervals.

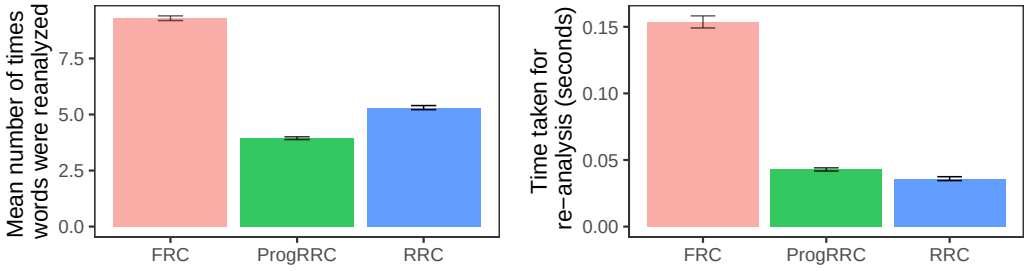
Model	Contrast	Estimate	CI	BF
Participial-Phrase (N = 1280 of 1280)	Intercept (Grand mean)	-9.26**	[-10.45, -8.25]	2.37e+21
	AMV vs all RCs	-1.64	[-4.24, 0.56]	1.38
	RRC vs. (ProgRRC & FRC)	4.17**	[2.30, 6.11]	1.00e+03
	ProgRRC vs. FRC	0.56	[-1.88, 3.08]	0.670
	Intercept (AMV)	-9.69**	[-11.57, -8.13]	1.63e+17
	RRC vs. AMV	3.71**	[1.86, 5.75]	931.14
	ProgRRC vs. AMV	-0.58	[-3.30, 1.84]	0.672
	FRC vs. AMV	-1.28	[-4.09, 1.22]	0.983
Whiz-Deletion (N = 1280 of 1280)	Intercept (Grand mean)	-7.24**	[-8.36, -6.16]	7.52e+09
	AMV vs all RCs	-3.52**	[-5.66, -1.91]	1.71e+03
	RRC vs. (ProgRRC & FRC)	1.39**	[0.57, 2.30]	57.15
	ProgRRC vs. FRC	-0.15×	[-0.146, 1.05]	0.308
	Intercept (AMV)	-8.74**	[-10.15, -7.43]	1.63e+17
	RRC vs. AMV	3.33**	[2.14, 4.56]	931.14
	ProgRRC vs. AMV	1.21	[-0.60, 2.83]	0.672
	FRC vs. AMV	1.76	[0.09, 3.29]	0.983
Participial-Phrase models (N = 130 of 1280)	Intercept (Grand mean)	-6.65**	[-7.87, -5.62]	1.32e+14
	AMV vs all RCs	-2.02	[-4.73, -0.24]	2.32
	RRC vs. (ProgRRC & FRC)	5.21**	[3.23, 7.24]	5.92e+03
	ProgRRC vs. FRC	0.59	[-1.94, 3.19]	0.693
	Intercept (AMV)	-7.43**	[-9.42, -5.80]	1.299e+14
	AMV vs RRC	4.92**	[3.00, 7.03]	1.85e+04
	AMV vs. ProgRRC	-0.94	[-3.96, 1.68]	0.833
	AMV vs. FRC	-1.66	[-4.62,1.10]	1.33
Whiz-Deletion models (N = 283 of 1280)	Intercept (Grand mean)	-5.04**	[-6.17, -4.02]	1.64e+07
	AMV vs all RCs	-3.72**	[-5.42, -2.35]	6.06e+03
	RRC vs. (ProgRRC & FRC)	1.14**	[0.37, 2.04]	15.99
	ProgRRC vs. FRC	-0.15×	[-1.42, 1.09]	0.294
	Intercept (AMV)	-6.95**	[-8.33, -5.76]	1.34e+11
	AMV vs RRC	3.66**	[2.66, 4.78]	1.37e+05
	AMV vs. ProgRRC	1.86*	[-0.04, 3.39]	3.79
	AMV vs. FRC	3.66**	[2.66, 4.78]	14.79

Table 2.9: Model estimates, 95% Credible Intervals and Bayes Factor estimates for data generated using the Whiz-Deletion and Participial-Phrase versions of the SPAWN model. The top half of the table indicates results form all 1280 model instances for each version and the bottom half indicates results from only the model instances that assigned at least one target prompt a RRC reading. The estimates are on the log odds scale and can be converted to probabilities using the following formula, where β is some estimate: $e^{\beta}/(1+e^{\beta})$. Bayes Factor estimates were computed using the Savage-Dickey method from the bayesfactor package in R (Makowski, Ben-Shachar, and Lüdecke, 2019a). Using the thresholds from Jeffreys (1939), we use * to indicate moderate evidence ** to indicate strong evidence for the alternative hypotheses. Similarly, we use × and ×× to indicate moderate and strong evidence for the null hypothesis.

condition was around 0.0005 for the Whiz-Deletion models and around 0.0003 for the Participial-Phrase models when all model instances were considered. As expected, these probabilities were higher when we considered only Passive-Models — around 0.04 in the Whiz-Deletion version and 0.07 in the Participial-Phrase version — but still very low. This predicts that even in the condition with the highest priming effect, and in only model instances that generated at least one reduced RC passive parse, less than 10% of target prompts received a reduced RC parse. If the proportion of participants who generate at least one reduced RC passive continuation is similarly low as is the the probability of these participants assigning the targets prompts a reduced RC passive, then we would expect the effects in our human experiment to also be very small and require many participants to detect.

In the Participial-Phrase model instances, as hypothesized in Equation 2.9, the probability of the target receiving a RRC parse was equivalent with the ProgRRC, FRC and AMV primes. In the Whiz-Deletion version, the probability of a RRC parse was highest in the RRC priming condition, followed by the ProgRRC and FRC conditions with equal probability and with the lowest probability in the AMV condition. This is contrary to our qualitative predictions in Equation 2.8 where we expected the probability of a RRC parse to be equal in all the RC conditions.

The hypothesis expressed in Equation 2.8 was based on the assumption that processing any of the RC primes would result in the same increase in the base-level activation of the *rc_noun* chunk required for a RRC parse because this chunk occurs in all of these types of sentences. However, posthoc analyses revealed that this assumption was not valid in our experimental setting: the RRC condition resulted in a greater mean increase in the base-level activation for the *rc_noun* chunk than



(a) **(b)**

Figure 2.3: Mean number of times re-analysis was triggered in each prime sentence (2.3a) and the mean amount of time taken (in seconds) to process each sentence (2.3b)

the other two conditions for 52% of all model instances.[10] This difference in base-level activations can account for the difference in predicted probability across the conditions.

Why was the mean increase in the base-level activation for the *rc_noun* chunk different across the different RC conditions? Recall that the base-level activation of a chunk depends not only on the number of times the model has encountered the chunk, but also on the amount of time that has passed since the model encountered the chunk (see Equation 2.3). Three factors can influence the amount of time between when the model is processing the subject noun in the target and when it retrieved the *rc_noun* chunk in the primes: the number of words in the prime sentences, the number of renalyses that get triggered when parsing the prime sentences and the amount of time each re-analysis takes. The number of words in the sentences were nearly identical across the structures with ProgRRC sentences having one additional

[10]Mean change in activation for the RRC condition was 0.148, and the mean change for the ProgRRC and FRC conditions were 0.12 and -0.10 respectively. Note, even though mean change for ProgRRCs was greater than FRCs, the predicted probabilities in both these conditions are equivalent (as evidenced by the Bayes Factors in Table 2.9) indicating that the predicted probability of RRC parses does not scale linearly with the mean change in activation for the *rc_noun* chunk after processing the primes.

word (i.e., 'being') and FRC sentences having two additional words (i.e., wh-phrase and finite auxiliary) compared to RRC sentences. This suggests that the differences in base-level activation across conditions is unlikely to be driven by just the number of words. We ran posthoc analyses to test whether the time taken to process prime sentences was greater in the FRC and ProgRRC conditions than in the RRC condition, and if so why.

When we compared FRC and RRC sentences, we discovered that the number of times re-analysis was triggered as well as the time taken for these re-analyses were much higher in FRC sentences than in RRC sentences (see Figure 2.3). This difference in the number of triggered re-analyses resulted from an interaction between the grammar we specified, the corpus frequencies in our training data and the mechanisms we assumed for dealing with null elements and re-analysis. We describe this interaction in the paragraph below before describing the cause for the difference between ProgRRC and RRC sentences.

In our training data, reduced passive and progressive RCs made up 31% of all RCs. Therefore, when the model is processing the subject noun in FRC sentences, there is a moderately high probability that the model incorrectly retrieves the *rc_noun* chunk followed by the *null_wh_subj* (i.e., the chunk associated with the null-wh phrase in RRCs and ProgRRCs). Once the model retrieves the *null_wh_subj* chunk, the model then has to retrieve one of the two auxiliary chunks and combine this with the previously retrieved *null_wh_subj* and *rc_noun* chunks (see Figure 2.1 and Algorithm 2). It is only after generating this combined state G_2 can the model proceed to process the overt wh-phrase "who" and the overt auxiliary "was" in FRC sentences. Given the grammar we specified, the parser is able to successfully combine the chunks

retrieved for the overt wh-phrase and auxiliary with G_2. A re-analysis only gets triggered when the model proceeds to process the verb in the embedded clause. Once the re-analysis gets triggered, the model backtracks step by step and tries all possible combination of chunks for the overt wh-phrase 'who', overt auxiliary 'was' and the null-auxiliary before re-processing the subject noun. This step-by-step backtracking process triggers many re-analysis processes, which in turn cause the model to process FRC primes more slowly than RRC primes.

When we compared ProgRRC with RRC sentences, we discovered that there were slightly fewer re-analyses triggered in ProgRRCs compared to RRCs (see Figure 2.3a). This difference resulted from the fact that the disambiguating word in ProgRRCs ("being") occurs earlier in the sentence than the disambiguating word in RRCs ("by"). Therefore, the parser can pursue incorrect parses for a longer duration, resulting in more re-analyses being triggered. Despite the fact that fewer re-analyses are triggered in ProgRRC sentences, the overall time taken to process these sentences is greater than the time taken for RRC sentences (see Figure 2.3b). The longer mean processing times for ProgRRC sentences was driven by the fact that the chunks associated with ProgRRCs (*progrrc_noun* and *null_progressive_aux*) have a lower base-level activation than the chunks associated with RRCs (*rrc_noun* and *null_finite_aux*) because ProgRRC sentences were much more infrequent than RRC sentences in our training data (see Table 2.8). Since the time taken to retrieve a chunk is inversely proportional to its activation (see Equation 2.5), the mean retrieval times were longer in ProgRRCs, in turn resulting in longer processing times.

2.4.3 Discussion

To generate quantitative predictions from the Whiz-Deletion and Participial-Phrase versions of the SPAWN models we first trained 1280 instances of both the Whiz-Deletion and Participial-Phrase versions on 100 random sentences each from a templatically generated dataset with 10000 sentences. Then, we tested all of these models on the experimental lists that we planned to test our participants on. All the hyperparameters of the model instances were consistent with prior ACT-R models. While the predicted patterns of priming aligned with our directional hypotheses in the previous section in the Participial-Phrase instances of SPAWN but not in the Whiz-Deletion instances of SPAWN. This deviation highlights the importance of generating quantitative predictions in models like SPAWN that are interpretable, but whose computation involves complex interactions. In the following section we describe the empirical experiment we conducted to test these predictions generated from the Whiz-Deletion and Participial-Phrase versions of SPAWN.

2.5 Which theoretical account best describes human sentence representations?

2.5.1 Methods

2.5.1.1 Participants

We recruited 769 participants via Prolific, a crowdsourcing platform. We only recruited participants whose first language was English, who were located in the US, and who did not participate in any of the pilot versions of this experiment. The median time taken to complete the experiment was 31.48 minutes and participants were

compensated with 8.35 USD. We excluded four participants who reported in our demographic survey that their first language was not English. The remaining 765 participants were included in our analyses.

2.5.1.2 Materials

When creating our stimuli, we picked 24 target verbs which can result in an MV/RR ambiguity (i.e., the past tense and the past Participial forms of these verbs are identical like in "examined"). We created four items per verb and four versions of each item. The four versions of one of the items for the verb "admired" is illustrated below.

(16) a. The singer admired by her fans sang beautifully (RRC).

b. The singer who was admired by her fans sang beautifully (FRC).

c. The singer being admired by her fans sang beautifully (ProgRRC).

d. The singer admired her fans and sang beautifully (AMV).

2.5.1.3 Design and Procedure

In any given trial in this experiment, participants were presented with either a complete sentence or a partial prompt and asked to memorize it. Once they memorized the sentence or prompt, participants progressed to a screen where they were required to re-type the sentence or prompt from memory. In trials where a prompt was presented, participants were additionally asked to complete the prompt. Participants were not able to progress to the next trial until they typed out the sentence or prompt perfectly, and in trials with a prompt, typed at least one additional word. They could go back to the screen with the sentence or prompt by clicking on a "Read prompt again" button. The experiment was divided into 24 prime-target blocks. Each block contained three

complete sentences (primes) followed by a partial prompt that was ambiguous between the MV and RR reading (target). There were 6 blocks assigned to each of the four priming conditions (RRC, FRC, ProgRRC and AMV). The sentences and prompt in each block had the same target verb. For example, one of the blocks that was assigned the FRC condition contained the following items:

(17) a. The singer who was admired by her fans sang beautifully.

b. The princess who was admired by the magician went into a trance.

c. The employee who was admired by the manager received a good evaluation.

d. The nurse admired ___.

Therefore in total, the experiment contained $6\times3\times4$=72 prime sentences and $6\times1\times4$ = 24 target prompts.

We randomly selected the order of the verbs to be presented and reversed this order to generate two random orders. For each random order we generated 16 counterbalanced lists, resulting in 32 lists in total. The lists were counterbalanced for prime type; for example, the first chunk in list 1 was assigned to RRC, in list 2 to FRC, in list 3 to ProgRRC and in list 4 to AMV. The lists were also counterbalanced for the order of the specific items in any given chunk; for example, item 1 was the first prime in list A, second prime in list B, third prime in list C and target in list D. Participants were randomly assigned to one of the 32 lists.

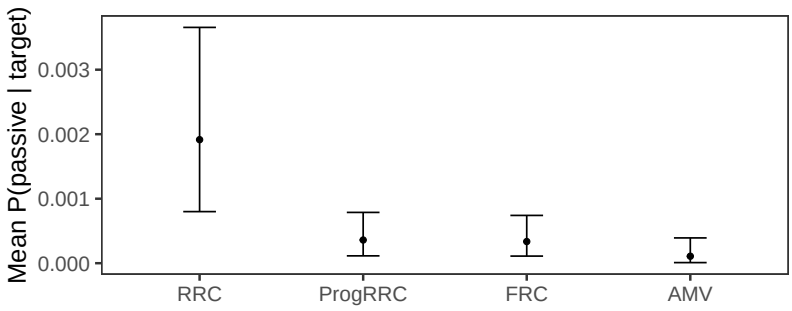


Figure 2.4: Probability of participants producing a passive continuation consistent with a RRC parse given an ambiguous target prompt. Plots were generated from the posterior samples of the mixed effects Bayesian Model 2 (described in 2.4.2.1). Error bars reflect 95% Credible Intervals.

Contrast	**Estimate**	**CI**	**BF**
Intercept (Grand mean)	-8.01	[-8.83, -7.27]	2.34e+24
AMV vs all RCs	-2.04*	[-4.09, -0.31]	6.45
RRC vs. (ProgRRC & FRC)	1.75**	[0.89, 2.65]	645.81
ProgRRC vs. FRC	0.07$^{\times}$	[-0.85, 0.77]	0.203
Intercept (AMV)	-8.83	[-10.24,-7.56]	9.94e+14
AMV vs RRC	2.63**	[1.28,4.05]	298.66
AMV vs. ProgRRC	0.35	[-1.23, 1.94]	0.45
AMV vs. FRC	0.67	[-0.92, 2.24]	0.57

Table 2.10: Model estimates, 95% Credible Intervals and Bayes Factor estimates for empirical data. The estimates are on the log odds scale and can be converted to probabilities using the following formula, where β is some estimate: $e^{\beta}/(1+e^{\beta})$. Bayes Factor estimates were computed using the Savage-Dickey method from the bayesfactor package in R (Makowski, Ben-Shachar, and Lüdecke, 2019a). Using the thresholds from Jeffreys (1939), we use * to indicate moderate evidence ** to indicate strong evidence for the alternative hypotheses. Similarly, we use $^{\times}$ and $^{\times\times}$ to indicate moderate and strong evidence for the null hypothesis.

2.5.2 Results

We fit the same Bayesian mixed effects models we used to analyze the predicted data generated from the Whiz-Deletion and Participial-Phrase versions of SPAWN (described in § 2.4.2.1). We list the model estimates and Bayes Factors in Table 2.10 and in Figure 2.4 we plot the probability of participants producing a passive continuation consistent with a RRC parse (i.e., P(passive | target)) under the different priming conditions. We computed these probabilities from the posterior distribution of the Helmert coded Bayesian model as described in 2.4.2.1.

As predicted by both the Whiz-Deletion and Participial-Phrase accounts, the probability of passive continuations was the highest in RRC primes. Numerically, this probability was higher with the ProgRRC and FRC primes than with AMV primes, which is consistent with the Whiz-Deletion, but not the Participial-Phrase account. However, Bayes Factor estimates indicated that there was insufficient evidence in our data to draw this conclusion. Like in our predicted data, this lack of sufficient evidence, despite having 765 participants, is a consequence of the empirical probability of passive continuations being very low — the mean empirical probability of passive continuations across all the conditions was only 0.0003. When the empirical probabilities are so low, finding strong evidence for differences between conditions would require extremely precise credible intervals, and therefore a very large number of participants.

The extremely low empirical probabilities of passive continuations is in turn results from the fact that 591 out of out 765 participants (i.e., 77.3%) never produced a passive continuation. Given the large proportion of participants who never produced a passive continuation in our experimental paradigm, merely increasing the number

of participants is unlikely to help us find strong evidence for the observed numerical differences between the AMV and the ProgRRC or FRC conditions. To estimate the impact that collecting more data can have on Bayes Factors for these comparisons, we generated new datasets with either 1024 participants or 1536 participants by re-sampling additional participants from our data and appending these participants to our original 765 participants. Repeating our analyses with these new datasets revealed that there was insufficient evidence to draw any conclusions about differences in the probability of passive continuations between the AMV and the ProgRRC or FRC conditions even with more than 1500 participants — the Bayes Factors were always between 0.43 and 1.10 (see Appendix for further details).

Why do most participants never generate a passive continuation? One possible explanation is that the low rate of passive continuations was a consequence of most participants not engaging with the task. The continuations that participants generated could have been a consequence of not how they parsed the ambiguous target prompt (as our linking hypothesis in Equation 2.1 assumes) but rather a consequence of strategy they might have used which helped them complete the task as soon as possible: the most effective strategy participants can adopt is to always produce continuations with a bare noun or simple noun phrase, irrespective of how they parsed the ambiguous target, since these continuations require participants to type only one or two additional words. If this explanation is accurate, then we would expect all (or most) of the 591 participants who never produced a passive continuation to also always produce continuations with a bare noun or a simple noun phrase. This was not the case, however: only 39 of the 591 participants always completed the prompt with only one or two additional words. We repeated our analyses by excluding these 39 participants

and found that the results were nearly identical (see Appendix). This suggests that a lack of engagement with the task was not driving the extremely low empirical probabilities of passive continuations.

An alternative explanation is that the low rate of passive continuations is direct consequence of the low frequency of reduced relative clauses. In the sentences that our participants encountered in their lives before participating in the experiment, ambiguous sequences like "the defendant examined" were almost always disambiguated in favor of the main verb parse. Therefore, in the experiment, it would be unsurprising if participants almost always assigned a main verb parse to the ambiguous target sequences. As discussed earlier, even in our SPAWN models, a majority of the model instances — 89.8% of models trained on the Participial-Phrase grammar 77.9% of models trained on the Whiz-Deletion grammar — never assigned a reduced relative clause parse to the ambiguous target sequences; only the model instances with a relatively high σ parameter (viz., the parameter that determines the amount of noise that gets added for any given trial) occasionally assigned reduced relative clause parses.

If some factor equivalent to a small σ parameter in our models prevented a group of participants from ever assigning a reduced relative clause parse to the ambiguous target sequences in our experimental setting, then it is not meaningful to ask how the probability of generating passive continuations changes is modulated by the different priming conditions in this group of participants. Since measuring this between-condition difference in the probability of passive continuations is crucial in evaluating the predictions between the two conditions, it would be more effective to break down the data analysis process to ask two separate questions. First, what proportion of participants produce at least one passive continuation? Second, of the participants

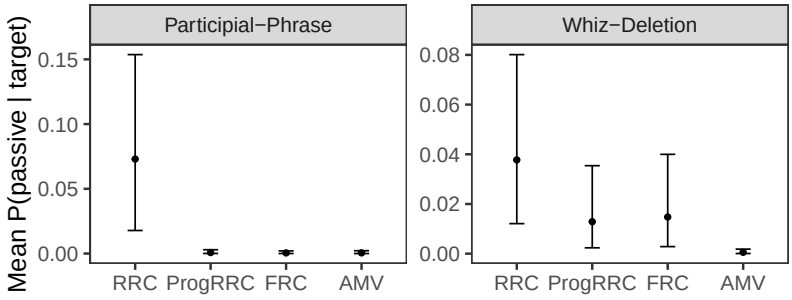
who produce at least one passive continuation, how is the probability of producing these continuations modulated by the different priming conditions?

The answer to the first question, as discussed earlier, is that 24% of out participants, i.e., 174 out of the 726 participants who engaged with the task[11] produced at least one passive continuation. To answer the second question, we repeated our analyses by only including participants who produced at least one passive response (or *Passive-Participants*). For a similar argument against aggregating data from different types of participants and an alternative method of analyzing such data, see Paape and Vasishth (2022).

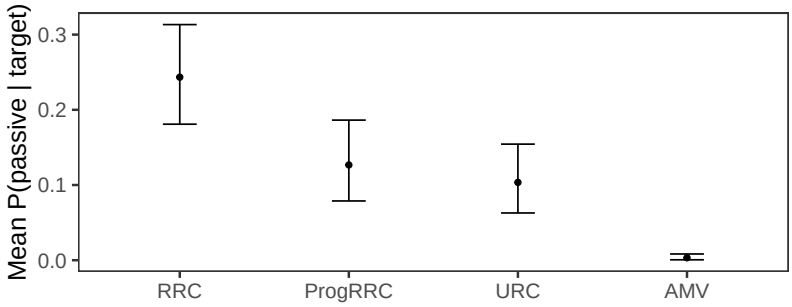
Measuring priming with the Passive-Participants Repeating our analyses with this subset of participants revealed that, as in our analyses with all participants, the probability of passive continuations was highest with the RRC primes. Crucially, Bayes Factor estimates revealed that there was strong evidence that the probability of passive continuations with ProgRRC and FRC primes was greater than the probability with AMV primes, and moderate evidence that the probability of passive continuations with ProgRRC primes was equivalent to that with FRC primes (see Table 2.11). This pattern of results aligns exactly with the qualitative pattern of predictions of the Whiz-Deletion version of the SPAWN model and not with the Participial-Phrase version (see Figures 2.2 and 2.5).

Qualitative vs. quantitative alignment The predicted probabilities in the Whiz-Deletion version greatly underestimated the empirical probabilities in all of the RC conditions (see Figure 2.5 and Table 2.11). This quantitative misalignment suggests

[11] Produced at least one continuation with more than two additional words.



(a) Predicted probabilities (Participial-Phrase N=377; Whiz-Deletion N=450)



(b) Empirical probabilities (N=174)

Figure 2.5: Probability of participants producing a passive continuation consistent with a RRC parse given an ambiguous target prompt. Plots were generated from the posterior samples of the mixed effects Bayesian Model 2 (described in 2.4.2.1). Error bars reflect 95% Credible Intervals.

Data	Contrast	Estimate	CI	BF
Human Participants (N=174 of 726)	Intercept (Grand mean)	-2.81**	[-3.32, -2.36]	8.22e+13
	AMV vs all RCs	-4.17**	[-5.71, -3.02]	9.72e+09
	RRC vs. (ProgRRC & FRC)	0.96**	[0.62, 1.31]	3.06e+04
	ProgRRC vs. FRC	0.21×	[-0.18, 0.60]	0.186
	Intercept (AMV)	-5.00	[-5.80,-4.30]	2.09e+18
	AMV vs RRC	3.86**	[3.15, 4.64]	8.54e+11
	AMV vs. ProgRRC	2.91**	[2.11, 3.75]	3.39e+05
	AMV vs. FRC	2.73**	[1.93,3.57]	6.57e+05
Whiz-Deletion models (N = 283 of 1280)	Intercept (Grand mean)	-5.04**	[-6.17, -4.02]	1.64e+07
	AMV vs all RCs	-3.72**	[-5.42, -2.35]	6.06e+03
	RRC vs. (ProgRRC & FRC)	1.14**	[0.37, 2.04]	15.99
	ProgRRC vs. FRC	-0.15×	[-1.42, 1.09]	0.294
	Intercept (AMV)	-6.95**	[-8.33, -5.76]	1.34e+11
	AMV vs RRC	3.66**	[2.66, 4.78]	1.37e+05
	AMV vs. ProgRRC	1.86*	[-0.04, 3.39]	3.79
	AMV vs. FRC	3.66**	[2.66, 4.78]	14.79
Participial-Phrase models (N = 130 of 1280)	Intercept (Grand mean)	-6.65**	[-7.87, -5.62]	1.32e+14
	AMV vs all RCs	-2.02	[-4.73, -0.24]	2.32
	RRC vs. (ProgRRC & FRC)	5.21**	[3.23, 7.24]	5.92e+03
	ProgRRC vs. FRC	0.59	[-1.94, 3.19]	0.693
	Intercept (AMV)	-7.43**	[-9.42, -5.80]	1.299e+14
	AMV vs RRC	4.92**	[3.00, 7.03]	1.85e+04
	AMV vs. ProgRRC	-0.94	[-3.96, 1.68]	0.833
	AMV vs. FRC	-1.66	[-4.62,1.10]	1.33

Table 2.11: Model estimates, 95% Credible Intervals and Bayes Factor estimates for predicted and empirical data, when considering only participants who produced at least one passive continuation and model instances which produced at least one reduced RC parse. Bayes Factor estimates were computed using the Savage-Dickey method from the bayestestR package in R (Makowski, Ben-Shachar, and Lüdecke, 2019b). Using the thresholds from Jeffreys (1939), we use * to indicate moderate evidence ** to indicate strong evidence for the alternative hypotheses. Similarly, we use × and ×× to indicate moderate and strong evidence for the null hypothesis.

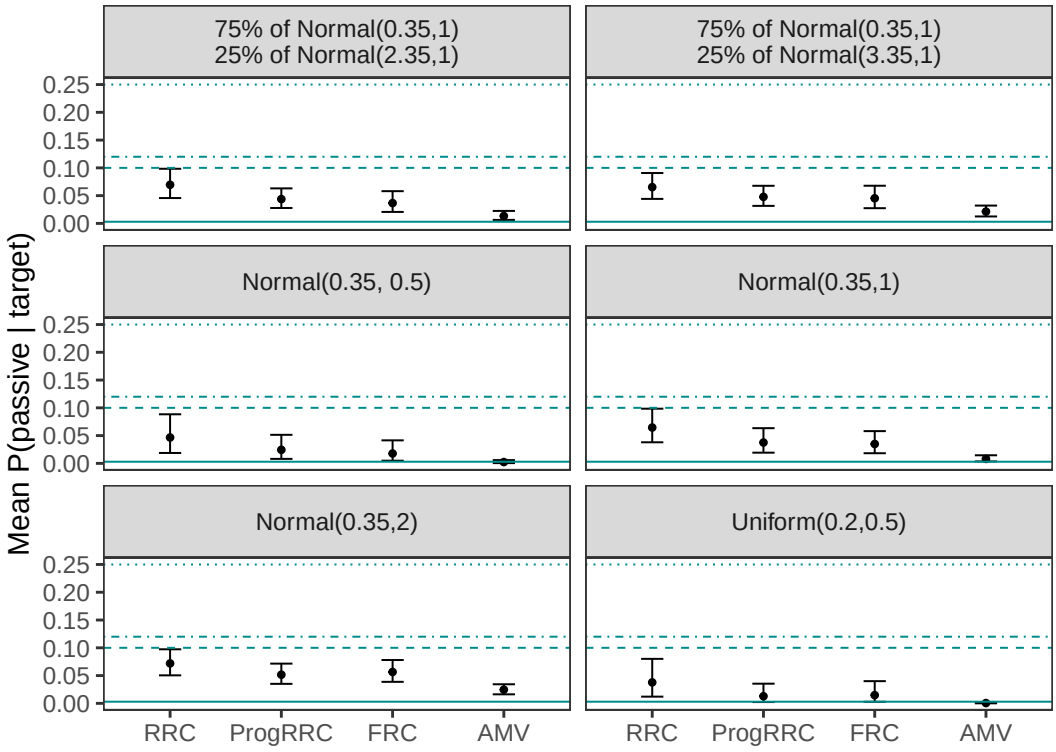


Figure 2.6: Predicted probability of a RRC parse given the target prompt under the Whiz-Deletion account. The facet labels indicate the distribution from which σ values were sampled for each of the 1280 model instances. The lines indicates the mean estimated probability of passive continuations in the human data under the different priming conditions. The dashed line corresponds to the AMV condition, the dashed line to the FRC condition, the dot-dashed line to the ProgRRC condition and the dotted line to the RRC condition. Error bars reflect 95% credible intervals.

that the hyperparameters and/or the parsing mechanism we specified in our SPAWN models was not accurate.

Of all the hyperparameters in our model (see 2.4.1.2), the distribution from which we sampled the Noise Parameter σ ($Uniform(0.2, 0.5)$) is the least motivated in prior work. Prior work specified this prior to capture the assumption that that values of the noise parameter ranging from 0.2 to 0.5 are most plausible (Vasishth and Engelmann, 2021). However, it is possible that for the phenomenon we are modeling in this work, higher values of σ might be required. Since reduced relative clauses are very infrequent, there is a large gap in the activation levels between the syntax chunks associated with a reduced RC parse (e.g., *rc_noun* and *Vt_pass*) and the chunks associated with the main verb parse (e.g., *noun* and *Vt_act*). Consequently, in order for the model to retrieve the *rc_noun* or *Vt_pass* chunks on any given trial, a higher amount of noise needs to be added when compared to chunks with a lower gap in activation levels compared to their competitors.

To test if the values of σ we sampled were the cause of the quantitative misalignment, we repeated our experiments by sampling σ values from different distributions. In these distributions, values ranging from 0.2 to 0.5 were still most probable (as assumed by Vasishth and Engelmann), but occasionally more extreme values could be sampled. These experiments revealed that merely changing the distribution that σ was sampled from was insufficient to get quantitative alignment with the empirical data. In order for the predicted and empirical probabilities to align, the predicted probabilities need to increase in the three RC conditions, but not the AMV condition. However, increasing the amount of noise in models' retrieval processes increased the probability of a reduced RC parse given the target prompt across all conditions (see Figure 2.6),

thus failing to correct the quantitative misalignment.

While it is possible that altering the other hyperparameters in our model can provide a better quantitative alignment with our empirical data, it is not desirable to do so because these hyperparameters have been successful in accounting for a wide range of psycholinguistic phenomena in prior work. Instead, in the following section we discuss how altering some of our simplifying assumptions about the parsing mechanism can result in better quantitative alignment, after we first summarize the experiment.

2.5.3 Discussion

To test the predictions generated by the SPAWN models under the Whiz-Deletion and Participial-Phrase accounts we collected data from 769 participants (765 included in our analyses). The overall empirical probability of participants completing ambiguous target prompts with continuations consistent with the reduced RC parse was very low — only 0.05% of all ambiguous target prompts were completed with passive continuations — which was a consequence of the fact that most of our participants ($N = 591$) never produced a passive continuation. Based on the length of the continuations that participants produced, we argued that the low proportion of participants who produced at least one passive continuation (Passive-Participants) was not a consequence of most participants not engaging with the task — if participants who never produced passive continuations were not engaged and were trying to complete the task as soon as possible, then most of these participants would only produce continuations with one or two additional words, which was not the case. Instead, we argued that the low proportion of Passive-Participants was a consequence of the infrequency of reduced

RCs; this argument was supported by the fact that only a small proportion of our model instances, under both the Whiz-Deletion and Participial-Phrase accounts, assigned at least one of the target prompts a passive continuation (Passive-Models).

Given the low proportion of Passive-Participants and Passive-Models, we argued that the effective way to evaluate the priming predictions under the two accounts was to adopt a two step analysis approach (cf., Paape and Vasishth 2022): first, compute the proportion of Passive-Participants and Passive-Models; second, within the subset of Passive-Participants and Passive-Models, measure the probability of passive continuations (or probability of a reduced RC parse in the case of models) given the target prompt under the different priming conditions. In order to conclude that our models under the Whiz-Deletion or Participial-Phrase accounts explain the empirical data, the following three factors have to be true: first, the proportion of Passive-Participants must be equivalent to the proportion of Passive-Models; second, the qualitative pattern of predicted probabilities of passive continuations under the different priming conditions must align with pattern of empirical probabilities; and finally, the magnitude of the differences in the empirical probability of passive continuations between the different priming conditions must align with the magnitude of differences in the predicted probabilities.

Our analyses revealed that, first, the proportion of Passive-Models in the models trained on both the Whiz-Deletion and Participial-Phrase grammars was comparable to the proportion of Passive-Participants. Second, the *qualitative* pattern of predicted probabilities of passive continuations aligned with the pattern of empirical probabilities in the Whiz-Deletion Passive-Models, but not the Participial-Phrase Passive-Models. And finally, even in the Whiz-Deletion Passive-Models, there was a *quantitative*

misalignment between the predicted and empirical probabilities, with the Passive-Models underestimating the probability of target prompt receiving a RRC parse in the three RC conditions. We argued that this quantitative misalignment was more likely to be driven by our assumptions about the parsing mechanism than the specific hyperparameters we chose. We now describe how our assumptions about the parsing mechanism can be altered in future work to improve the quantitative alignment.

Changing assumptions about the parsing mechanism to improve quantitative alignment in future work One explanation for why our SPAWN models underestimated the probability of a RRC parse given the target sequence is that our simplifying assumption about strictly serial parsing does not account for all the sources of facilitation that can occur due to priming. Under the strictly serial parsing assumption, the probability of a RRC parse given the target sequence is determined entirely by the probability that the model retrieves the chunk *rc_noun* when processing the subject noun (e.g., "defendant" in the sequence "the defendant examined"). Therefore, under the Whiz-Deletion account, any change in the probability of a RRC parse given the target sequence after processing RC prime sentences is solely driven by a change in the base-level activation of the *rc_noun* chunk after processing the prime sentences.

However, prior psycholinguistic work has revealed that the probability of a reduced RC parse is also impacted by the probability that the model retrieves the chunk *Vt_pass* when processing the ambiguous verb (e.g., *examined* in the sequence 'the defendant examined"). The evidence for this comes from experiments which demonstrate that verb-repetition increases the magnitude of the priming effect (for a review, see Mahowald et al. 2016). In our experiments, since there was verb repetition between the prime and target, the lexical activation from the verb to the *Vt_pass* chunk increased

along with the base-level activation of this chunk, thus resulting in a greater increase in the probability of this chunk being retrieved when compared to experiments with no verb-repetition. However, since the probability of the *Vt_pass* chunk being retrieved did not impact the probability of the target receiving a RRC parse in our models (as discussed in the previous paragraph), this boost in activation due to verb repetition was not reflected in our models' predictions.

The probability of the *Vt_pass* chunk being retrieved can impact the probability of the target receiving a RRC parse if we assume a parallel parsing mechanism. In a parallel parsing mechanism, the parser maintains many, if not all, of the incremental parses that are possible given a sequence. After the parallel version of our model completes parsing a target sequence like "the defendant examined", both the reduced RC and the main verb parses are available to the model. Therefore, the probability with which the model assigns a RRC reading to the target sequence will be impacted by the total probability of the RRC parse given the target sequence, which is influenced by both the probability of subject noun being associated with the *rc_noun* chunk and the probability of the verb being associated with the *Vt_pass*.

Thus, unlike our current SPAWN models, the predictions from SPAWN models with a parallel parsing mechanism will be sensitive to verb repetition effects. Consequently, a version of our SPAWN models with a parallel parsing mechanism will likely predict larger priming effects, as was observed in the empirical data. This observation that a parallel parsing mechanism is necessary to account for the full pattern of empirical data aligns with other work from Boston et al. (2011), who demonstrated that in sentence comprehension models that derive predictions about eye-tracking data, a parallel parsing mechanism made "the difference between empirical adequacy

and empirical inadequacy".

Adding subject noun repetition to find stronger priming effects in future work As described in the previous paragraph, the probability of some chunk c being retrieved when a word w is being processed depends on both the base-level activation of c as well as the activation that c receives from w. Therefore, under the SPAWN model trained with the Whiz-Deletion grammar, repeating the subject noun across the prime and target will increase the lexical activation from the noun to the *rc_noun* chunk, which in turn will increase the probability of this chunk being retrieved, and consequently resulting in a RRC parse of the target prompt. The model of priming proposed by Reitter, Keller, and Moore (2011) also predicts facilitation by repeating the subject noun. Therefore, in order to increase the magnitude of priming and consequently the power of the experiment, future work using our proposed experimental paradigm can also repeat the subject noun between the prime and targets.

2.6 General discussion

In this chapter, we proposed a method of characterizing the incremental structures that comprehenders build when processing sentences in real time. Our proposed method draws on hypotheses about the abstract structure of sentences from generative syntax, and converts these hypotheses into testable quantitative behavioral predictions. To generate these behavioral predictions, we developed a model of serial parsing: SPAWN. The incremental structures that are built in SPAWN are influenced by the computational principles within the ACT-R framework (Anderson et al., 2004), which is a general cognitive architecture designed to account for a wide range of cognitive

phenomena.

As a case study, we used SPAWN to study the incremental structures that comprehenders build when reading sentences with reduced RCs. We introduced competing hypotheses in generative syntax about the underlying structure of reduced RCs: the Whiz-Deletion account which argues that the structure of all RCs, whether reduced or not, contains a CP node; and the Participial-Phrase account which argues that only the structure of full, but not reduced RCs, contains a CP node. We evaluated which of these two hypotheses better characterizes the structures that people construct using a four step approach. First, we specified two grammars using the CCG formalism, one consistent with the representational assumptions of Participial-Phrase account and another consistent with the assumptions of the Whiz-Deletion account. Second, we trained separate SPAWN models on the two grammars. Next, using these trained models, we generated behavioral predictions for a comprehension-to-production experiment under the Whiz-Deletion and Participial-Phrase accounts. Finally, we tested these predictions by running a large-scaled behavioral experiment ($N = 765$).

In every trial of our comprehension-to-production experiment, human or SPAWN participants were presented with three prime sentences followed by a target prompt like "The defendant examined" which was ambiguous between a main verb reading (i.e., the defendant examined someone) and a reduced RC (i.e., the defendant was examined by someone). The question of interest was how the probability with which participants assigned a reduced RC parse to the ambiguous target prompt changed as a function of the structure of the prime sentences — i.e., the extent to which different types of sentences with and without RCs *primed* a reduced RC reading of ambiguous prompts. In our SPAWN participants, we inferred the parse assigned to the target

prompt using the syntactic categories assigned to the words in the prompt, whereas, in our human participants we inferred the parse based on how these participants completed the target prompt.

Summary of the patterns of behavior in SPAWN participants In our Whiz-Deletion grammar, the subject nouns in all the three types of relative clause primes we considered (RRC, ProgRRC and FRC) were assigned the NP/CP category, thus capturing the assumption under the Whiz-Deletion account that all relative clauses contain a CP node. In the models trained on this grammar, the probability of the target prompt receiving a reduced RC parse depended on the probability with which these models retrieved the NP/CP category when processing the subject noun in the prompt. Consequently, in these models, the probability of the target prompt receiving a RRC parse was higher in the RC conditions when compared to the AMV condition: the models always retrieved the NP/CP category when processing the RC primes but not the AMV primes, and the frequency with which a category was retrieved in the past is proportional to the probability of the category being retrieved (see Equation 2.3) . Additionally, in these models, the probability of the target prompt receiving a RRC parse was higher in the RRC condition compared to the other two RC conditions. This difference was not a consequence of the grammar, but rather a consequence of the corpus frequencies and the parsing mechanism we assumed (see § 2.4.2.2 for a detailed discussion).

In contrast, in our Participial-Phrase grammar, the subject nouns in the RRC, FRC and ProgRRC prime conditions were all assigned different syntactic categories — NP/VoiceP, NP/CP and NP/(VoiceP/ProgP) respectively — thus capturing the intuition that reduced passive and progressive RCs do not contain a CP node. In the models

trained on this grammar, the probability of the target prompt receiving a reduced RC parse depended on the probability with which these models retrieved the NP/VoiceP category when processing the subject noun in the prompt. As a consequence of this grammar, in these models, the probability of the target prompt receiving the reduced RC parse was highest in the RRC condition and equivalent in all the other three conditions because the models retrieved the NP/VoiceP category only when processing the RRC primes. Therefore the probability of the target prompt receiving a reduced RC parse was highest in the RRC condition and equivalent in all of the other conditions.

Insights from qualitative alignment between human and SPAWN participants

The behavioral patterns from our human experiment qualitatively aligned with the patterns from the SPAWN models trained on the Whiz-Deletion grammar: the probability of target prompts being assigned a reduced RC parse was highest in the RRC prime condition, lowest in the AMV condition and equivalent in the other two RC conditions. Since the SPAWN models are interpretable — i.e., we understand what incremental structures they produce and why — this qualitative alignment can provide insight into the factors that can influence the incremental structures that human comprehenders construct when processing sentences with reduced RCs.

The pattern of behaviour in human and SPAWN participants can be broken down into two parts. First, the probability of target prompts receiving a reduced RC parse is greater in the conditions with RC primes compared to the AMV prime condition without RCs. As discussed above, this behaviour emerges in the SPAWN models as a consequence of the representational assumption under the Whiz-Deletion account that all RCs share some abstract structure — specifically a CP node. Since human

participants also display this pattern of behavior, we can infer that this behavior is likely driven by the fact that there are some shared properties between the structures that human comprehenders construct as they are processing sentences with different types of RCs.

The second pattern of behavior is that the probability of target prompts receiving a reduced RC parse is greater in the RRC prime condition compared to the ProgRRC and FRC prime condition. As discussed above (and in § 2.4.2.2), this behavior emerges in the SPAWN models as a consequence of the corpus frequencies we used and the specific mechanisms we proposed to process null elements and recover from incorrect parsing decisions. Since human participants also display this pattern of behavior, we can infer that this behaviour is more likely to be a consequence of comprehenders' parsing mechanism than their grammar — even if the specific re-analysis or parsing mechanisms that human comprehenders use are different from the ones we proposed. If this inference is accurate, then these predictions can serve as data points that constrain future work focused on building more cognitively plausible models of parsing and re-analysis.

Insights from the lack of quantitative alignment between human and SPAWN participants While the pattern of behavior in the SPAWN models trained with the Whiz-Deletion grammar aligned *qualitatively* with the behaviour of human participants, the difference in the probability of a reduced RC parse given the target prompt between the priming conditions was greater in the human participants than in the SPAWN models. In § 2.5.3, we argued that this quantitative misalignment was likely a consequence of our simplifying assumption that the parsing mechanism is strictly serial. This conclusion is consistent with other modeling work (cf. Boston et al. 2011)

and highlights the necessity of developing a parallel version of the SPAWN model — i.e., a *Parallel* Parsing in ACT-R With Null elements (PPAWN) model — in future work.

Can this work inform generative syntactic theory? On one end of the spectrum, a classical view of syntax (Chomsky, 1965) argues that psycholinguistic evidence is irrelevant for theory building in generative syntax because this evidence is thought to reflect *performance* (i.e., factors that influence how humans use their linguistic knowledge in real time) and not *competence* (i.e., the linguistic knowledge itself). For instance, as described in § 2.2, Harwood (2018) proposed a theory of reduced RCs, which he argued explained the inflectional restrictions in English better than existing theories. A syntactician might challenge Harwood's proposal on the grounds that the representations that the Participial-Phrase assumes are adhoc and not supported by convicing evidence. It is beyond the scope of this work to evaluate the evidence that supports the Participial-Phrase account. The relevant point here is that under a classical view of syntax, competing hypotheses in theoretical syntax (such as the Whiz-Deletion and Participial-Phrase hypotheses) should be evaluated on their ability to most elegantly explain a wide range of acceptability judgment data, ideally across many of the world's languages, rather than on their ability to account for psycholinguistic evidence.

On the other end of the spectrum, Branigan and Pickering (2017) argue that theory building in syntax should "end the current reliance on acceptability judgments" and instead rely on psycholinguistic evidence. Therefore, under the view proposed by Branigan and Pickering, Whiz-Deletion is a better hypothesis than Participial-Phrase because the predictions from the Whiz-Deletion hypothesis better aligned

with our human behavioral data. In an open peer commentary of this work, several authors challenged this view and argued that psycholinguistic evidence cannot replace acceptability judgments: acceptability judgments are better suited to study some phenomena that are of interest to syntactic theory (Ambridge, 2017; Gaston, Huang, and Phillips, 2017) because psycholinguistic experiments might not be sensitive to the relevant distinctions (Adger, 2017; Ruiter and Ruiter, 2017; Koring and Reuland, 2017; Martin, Huetting, and Nieuwland, 2017; Rees and Bott, 2017; Ryskin and Brown-Schmidt, 2017).

We adopt a less extreme view and argue that data from psycholinguistic experiments should be one of factors that is used to evaluate competing hypotheses in syntactic theory alongside acceptability judgments. This view aligns with work focused on integrating modern syntactic theory with psycholinguistic evidence (Franck et al., 2006; Kobele, Gerth, and Hale, 2013; Graf, Monette, and Zhang, 2017; De Santo, 2021). Under this view, a theory of reduced relative clauses should be able to account for both the inflectional restrictions in English as well as the priming data in our behavioral experiment which suggests that the structures that comprehenders build when processing relative clauses share some abstract properties.

Future work In this work, we used the SPAWN models to generate *offline* predictions about the continuations to these target prompts participants were expected to produce under the different priming conditions. Our model can also be used to generate *online* predictions about the amount of time participants are expected to spend reading specific words in sentences (see Equation 2.5). These predictions can be evaluated using existing data from self-paced reading or eye-tracking experiments, such as the recently introduced large-scaled benchmark called the Syntactic

Ambiguity Processing (SAP) benchmark (Huang et al., 2022). The SAP benchmark includes word-by-word reading time data from 2000 participants who read a wide range of temporarily ambiguous sentences, including reduced relative clauses. The wide range of phenomena included in this benchmark combined with the large number of participants, makes it possible to compute very precise sentence-level reading time estimates for a broad range of phenomena. These precise estimates make it possible to evaluate the extent to which the specific mechanisms we proposed in this work can generalize to phenomena beyond reduced relative clause parsing. Since the computational mechanisms of SPAWN are transparent, cases in which the model predictions do not align with empirical data can suggest concrete ways of fine-tuning the parsing and re-analysis mechanism.

Generating predictions for the sentences in the SAP Benchmark from SPAWN models is relatively straightforward: it only requires adding additional syntactic categories and lexical items to the existing grammar(s). However, it is highly unlikely that the predictions from the current version of the SPAWN models will align with the item-level data in the SAP Benchmark for several reasons. First, as discussed earlier, our approximate non-restrictive implementation of type-raising is not suitable for generating timing predictions. Additionally, the strictly serial parsing assumption we made is not accurate. Therefore it is quite likely that a parallel version of SPAWN with an adequately restrictive type-raising mechanism is required to capture the whole range of effects in the benchmark. Second, our current method for generating the training data set for the SPAWN models does not take into account factors like lexical frequency, verb-bias or plausibility, all of which are very likely to result in sentence-level differences in reading times. Therefore, it is necessary to include these factors

in the templates used to create the training datasets, or train the model on more naturalistic datasets such as the CCGBank (Hockenmaier and Steedman, 2007).

2.7 Conclusion

In this chapter, we proposed a novel model of parsing using ACT-R — SPAWN — which can be used to convert hypotheses about comprehenders' incremental structural representations into testable behavioural predictions. As a case study, we used this model to study the incremental structural representations that comprehenders construct when reading temporarily ambiguous sentences with reduced relative clauses. First, we generated behavioural predictions from two competing hypotheses about the underlying structure of sentences with reduced relative clauses: the Whiz-Deletion account and the Participial-Phrase account. Then, we tested these predictions using a large-scaled web-based comprehension-to-production priming experiment. We demonstrated that the empirical data qualitatively aligned with the predictions of the Whiz-Deletion but not the Participial-Phrase hypothesis. We identified that the predictions that the models generated were influenced not only by the grammar we assumed under the Whiz-Deletion account, but also by the specific parsing mechanisms we assumed. Thus, based on the *qualitative alignment* between the predicted and empirical data, we were able to infer which parts of participants' behaviour were more likely to be driven by their underlying grammar than the parsing mechanism and vice versa. We also demonstrated that the empirical data underestimated the magnitude of effects in certain conditions, which we attributed to the serial parsing simplifying assumption we made in our model. Thus, based on the *quantitative misalignment* between the predicted and empirical data, we were able to identify concrete ways of

improving our model in future work.

2.8 Acknowledgments

This work was supported by an American Psychological Association Dissertation Award. We would like to thank Aniello de Santo, Will Merill, Shravan Vasishth and Suhas Arehalli for insightful discussions which helped shape this work. We would also like to thank members of the Computation and Psycholinguistics lab and the audience of the HSP 2022 Satellite at UMD for valuable feedback on early stages of this work.

2.9 Appendix

2.9.1 Assumptions underlying the specified syntax trees

The syntax trees in Figures 2.1 and 2.2 are constructed based on the following assumptions made by (Harwood, 2018).

- A head external analysis of RCs in which the head NP is generated outside the relative clause CP. The wh-phrase co-indexed with the head NP is base-generated in-situ and moves to the specifier of CP if relevant (Quine, 1960).
- What-you-see-is-what-you-get (WYSIWYG): if a sentence does not have a certain inflection, then the phrase associated with the inflection is absent from the tree. For example, ProgP is absent from sentences without the progressive inflection. Similarly VoiceP is absent from active sentences because active voice is assumed to be unmarked.
- All auxiliary verbs raise to inflectional heads to undergo abstract feature checking of inflection. Lexical verbs, on the other hand, do not raise. Rather, they undergo an Agree relation with higher inflectional heads and the features are checked in-situ.
- If there are no inflectional heads above a certain phrase (e.g., above VoiceP or ProgP), then there is no requirement for the syntax to merge into a higher vP shell headed by BE (e.g., vP in the case of VoiceP and vPProg in the case of ProgP). So in RRC sentences, the Clause Internal Phase (CIP) is headed by a VoiceP and in ProgRRC sentences the CIP is headed by ProgP.

- The aspectual hierarchy in Standard English is: Tense > Perfect Aspect > Progressive Aspect > Voice > Verb.

2.9.2 Results with different distributions for sampling σ

SD distribution	Model	Contrast	Estimate	CI	BF
$Normal(0.35, 0.5)$	Participial-Phrase (221 models)	Intercept (Grand mean)	-5.17**	[-6.00, -4.50]	7.54e+17
		AMV vs all RCs	-1.20	[-3.05, 0.30]	1.130
		RRC vs. (ProgRRC & FRC)	4.37**	[3.27, 5.71]	1.75e+07
		ProgRRC vs. FRC	-0.32	[-2.07, 1.31]	0.431
	Whiz-Deletion (355 models)	Intercept (Grand mean)	-4.32**	[-5.22, -3.51]	2.14e+07
		AMV vs all RCs	-2.55**	[-3.73, -1.66]	1.71e+04
		RRC vs. (ProgRRC & FRC)	0.88**	[0.35, 1.48]	23.15
		ProgRRC vs. FRC	0.36×	[-0.49, 1.26]	0.282
$Normal(0.35, 1)$	Participial-Phrase (377 models)	Intercept (Grand mean)	-3.52**	[-3.88, -3.18]	1.27e+24
		AMV vs all RCs	-0.89**	[-1.55, -0.34]	43.31
		RRC vs. (ProgRRC & FRC)	1.84**	[1.31, 2.37]	1.07e+06
		ProgRRC vs. FRC	0.21×	[-0.34, 0.79]	0.181
	Whiz-Deletion (450 models)	Intercept (Grand mean)	-3.55**	[-3.07, -4.07]	1.32e+11
		AMV vs all RCs	-1.75**	[-2.44, -1.15]	1.37e+04
		RRC vs. (ProgRRC & FRC)	0.62**	[0.31, 0.96]	53.08
		ProgRRC vs. FRC	0.07×	[-0.47, 0.69]	0.138
$Normal(0.35, 2)$	Participial-Phrase (657 models)	Intercept (Grand mean)	-2.53**	[-2.81, -2.27]	1.81e+15
		AMV vs all RCs	-0.55**	[-0.79, -0.33]	2.83e+03
		RRC vs. (ProgRRC & FRC)	0.89**	[0.71, 1.08]	3.19e+06
		ProgRRC vs. FRC	-0.11××	[-0.34, 0.12]	0.091
	Whiz-Deletion (589 models)	Intercept (Grand mean)	-3.01**	[-3.32, -2.70]	5.24e+13
		AMV vs all RCs	-0.92 **	[-1.27, -0.60]	7.57e+03
		RRC vs. (ProgRRC & FRC)	0.31	[0.06, 0.54]	1.45
		ProgRRC vs. FRC	-0.10××	[-0.38, -0.10]	0.089
Mixture of 1. $Normal(0.35, 1)$ 2. $Normal(2.35, 1)$ 75% of 1; 25% of 2	Participial-Phrase (535 models)	Intercept (Grand mean)	-2.94**	[-3.26, -2.63]	4.92e+13
		AMV vs all RCs	-0.54**	[-0.90, -0.24]	26.21
		RRC vs. (ProgRRC & FRC)	1.23**	[0.99, 1.46]	1.96e+07
		ProgRRC vs. FRC	-0.22×	[-0.58, 0.16]	0.205
	Whiz-Deletion (520 models)	Intercept (Grand mean)	-3.34**	[-3.75, -2.96]	4.29e+13
		AMV vs all RCs	-1.35**	[-2.02, -0.79]	669.71
		RRC vs. (ProgRRC & FRC)	0.59**	[0.37, 0.83]	1.27e+03
		ProgRRC vs. FRC	0.20×	[-0.27, 0.68]	0.166
Mixture of 1. $Normal(0.35, 1)$ 2. $Normal(3.35, 1)$ 75% of 1; 25% of 2	Participial-Phrase (589 models)	Intercept (Grand mean)	-2.54**	[-2.85, -2.25]	3.34e+15
		AMV vs all RCs	-0.49**	[-0.75, -0.27]	199.84
		RRC vs. (ProgRRC & FRC)	0.91**	[0.71, 1.10]	2.16e+06
		ProgRRC vs. FRC	-0.27	[-0.55, 0.01]	0.507
	Whiz-Deletion (568 models)	Intercept (Grand mean)	-3.15**	[-3.53, -2.79]	9.57e+13
		AMV vs all RCs	-0.93**	[-1.33, -0.61]	4.64e+03
		RRC vs. (ProgRRC & FRC)	0.37**	[0.16, 0.57]	10.44
		ProgRRC vs. FRC	0.06×	[-0.37, 0.51]	0.112

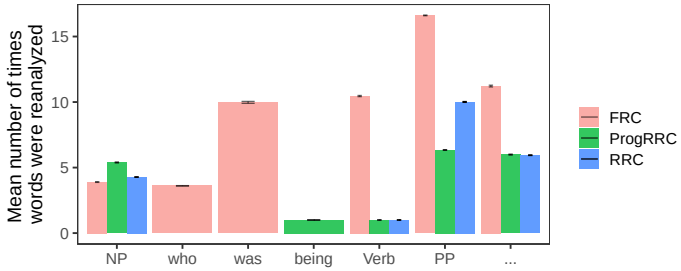
Table 2.12: Effect of the distribution that σ was sampled from on predictions from the Whiz-Deletion and Participial-Phrase versions of the SPAWN model. Using the thresholds from Jeffreys (1939), we use * and × to indicate moderate evidence ** and ×× to indicate strong evidence for the alternative and null hypotheses respectively.

2.9.3 Estimating changes in Bayes Factors with more data

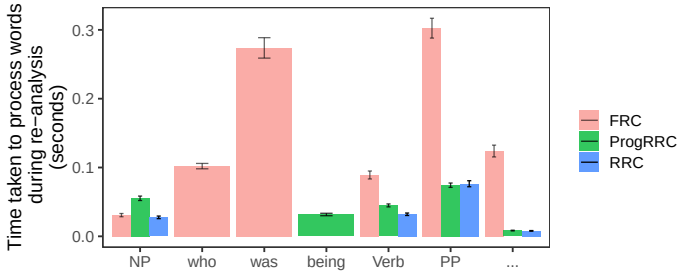
# Participants	Contrast	Mean estimate	Range estimate	Mean BF	Range BF
1024	Intercept (AMV)	-9.15	[-9.52, -8.76]	2.89e+16	[2.35e+15, 9.77e+16]
(259 resampled)	RRC vs. AMV	2.69	[2.31, 2.97]	8.92e+02	[1.10e+02, 2.80e+03]
	ProgRRC vs. AMV	-0.16	[-0.23, 0.18]	0.43	[0.42, 0.44]
	FRC vs. AMV	0.33	[-0.24, 0.69]	0.47	[0.40, 0.58]
1536	Intercept (AMV)	-9.98	[-10.44, -9.66]	2.52e+17	[1.24e+14, 1.23e+18]
(771 resampled)	RRC vs. AMV	3.07	[2.89, 3.34]	2.58e+03	[6.87e+02, 4.61e+03]
	ProgRRC vs. AMV	0.12	[-0.21, 0.59]	0.46	[0.43, 0.53]
	FRC vs. AMV	0.77	[0.37, 1.15]	0.71	[0.48, 1.10]

Table 2.13: Model estimates and Bayes Factors from the baseline coded Bayesian Mixed Effects Model for simulated datasets with 1024 and 1536 participants. The simulated datasets were constructed by resampling novel participants and adding them to the original dataset. We generated five datasets for each dataset size. The table lists the mean of each estimate and the corresponding Bayes Factor averaged across all five datasets, as well as the range of these estimates and Bayes Factors across these datasets.

2.9.4 Number of triggered re-analyses and the time taken to process prime sentences broken down by region



(a)



(b)

Figure 2.7: Mean number of times re-analysis was triggered in each prime sentence (2.7a) and the mean amount of time taken (in seconds) to process each sentence (2.7b)

2.9.5 Stimuli

Below is the list of all RRC sentences used in the experiment. The FRC, ProgRRC and AMV sentences were only minimally different from the RRC sentences as illustrated below for the first sentence. The target prompts were generated by taking the first three words of the RRC sentences.

(18) The prince accompanied by the duke arrived at the palace. (RRC)

(19) The prince who was accompanied by the duke arrived at the palace. (FRC)

(20) The prince being accompanied by the duke arrived at the palace. (ProgRRC)

(21) The prince accompanied the duke and arrived at the palace. (AMV)

l The prince accompanied by the duke arrived at the palace.

The children accompanied by their guardian skipped to the park.

The toddler accompanied by her parents went to the store.

The builder accompanied by the architect visited the construction site.

The singer admired by her fans sang beautifully.

The princess admired by the magician went into a trance.

The employee admired by the manager received a good evaluation.

The nurse admired by the doctor worked diligently.

The student approached by the teacher passed the exam.

The clerk approached by his manager submitted the report.

The spy approached by the agent searched for some answers.

The pharmacist approached by the patient recommended the blue pills.

The troops attacked by the terrorists suffered some heavy losses.

The lion attacked by the hunter staggered into the forest.

The senator attacked by the assassin went to the hospital.

The analyst attacked by his client apologized.

The politician betrayed by the party denied the allegations.

The king betrayed by the genie arrested the innocent woman.

The businessman betrayed by his partner lost a lot of money.

The executive betrayed by her accountant committed the famous tax fraud.

The policeman captured by the kidnapper divulged a lot of information.

The witch captured by the peasant cackled gleefully.

The suspect captured by the investigator looked terrified.

The prisoners captured by the colonel escaped at night.

The dog chased by the boy played with the ball.

The monkey chased by a hatter stole his hats.

The thief chased by the officer stopped suddenly.

The dentist chased by her son threatened playfully to extract his teeth.

The captain congratulated by the team jumped excitedly.

The actress congratulated by the director talked to the media happily.

The governor congratulated by the president made a public statement.

The valedictorian congratulated by her family beamed joyfully.

The widow consoled by her family felt better eventually.

The teenager consoled by the friend went back home sullenly.

The woman consoled by her daughter smiled wistfully.

The athlete consoled by his coach promised to put in more effort.

The boy described by the lady glowed joyfully.

The secretary described by the visitor waited patiently.

The chemist described by the writer was interested in philosophy.

The physicist described by the journalist published many seminal articles.

The gymnast encouraged by the coach participated in the contest.

The musician encouraged by the crowd performed a breathtaking piece.

The teacher encouraged by her students brought some donuts.

The hairdresser encouraged by her customer tried a new hairstyle.

The defendant examined by the lawyer was unreliable.

The engineer examined by a licensor passed the test.

The industrialist examined by the auditor hid the money.

The patient examined by the hygienist looked uncomfortable.

The thief identified by the victim ran away in fear.

The warden identified by the prisoner cracked the walnuts enthusiastically.

The student identified by the researcher proposed an interesting project.

The electrician identified by the company was competent.

The landlord loved by his tenants was considerate.

The comedian loved by the audience impersonated a famous celebrity.

The kid loved by her parents slept peacefully.

The janitor loved by the employees was cheerful.

The cashier paid by the customer muttered something inaudibly.

The organization paid by the government expanded its mission.

The broker paid by the company signed the contract.

The entrepreneur paid by the industrialist founded a successful company.

The man recognized by the spy took off down the street.

The actor recognized by the reporter waved enthusiastically.

The poet recognized by the writer smiled.

The driver recognized by the passenger grinned cheerfully.

The priest recorded by the mob preached non-violence.

The management recorded by the employee talked about privacy.

The dancer recorded by his trainer practiced energetically.

The influencer recorded by her fans started trending rapidly.

The girl scratched by the cat enjoyed the sunny afternoon.

The toddler scratched by the table exclaimed in surprise.

The dog scratched by a stranger barked loudly.

The diamond scratched by the machine remained flawless.

The applicant selected by the company gained more information.

The professor selected by the university made plans to move.

The chef selected by the restaurant advertised his signature dish.

The developer selected by the team solved the problem efficiently.

The culprit sketched by the investigator escaped.

The apprentice sketched by the artist rolled up the canvas.

The painter sketched by her student looked happy.

The robot sketched by the illustrator gained popularity rapidly.

The soldier studied by the general fought valiantly.

The writer studied by the scholar wrote a radical book.

The apprentice studied by the blacksmith forged a shield.

The scholars studied by the psychologist were attentive.

The officer envied by the assistant quit the job.

The principal envied by the administration complained incessantly.

The man envied by his sister bought a new car.

The plumber envied by the carpenter was well-known.

The speaker introduced by the chairman charmed everyone.

The producer introduced by the association won many awards.

The elf introduced by the magician danced happily.

therapist introduced by the doctor listened carefully.

The protesters challenged by the authority marched on the streets.

The scientist challenged by the journalist demonstrated the invention.

The dragon challenged by the wizard roared loudly.

The programmer challenged by the supervisor accepted her mistake.

The prince accompanied by the duke arrived at the palace.

The children accompanied by their guardian skipped to the park.

The toddler accompanied by her parents went to the store.

The builder accompanied by the architect visited the construction site.

The singer admired by her fans sang beautifully.

The princess admired by the magician went into a trance.

The employee admired by the manager received a good evaluation.

The nurse admired by the doctor worked diligently.

The student approached by the teacher passed the exam.

The clerk approached by his manager submitted the report.

The spy approached by the agent searched for some answers.

The pharmacist approached by the patient recommended the blue pills.

The troops attacked by the terrorists suffered some heavy losses.

The lion attacked by the hunter staggered into the forest.

The senator attacked by the assassin went to the hospital.

The analyst attacked by his client apologized.

The politician betrayed by the party denied the allegations.

The king betrayed by the genie arrested the innocent woman.

The businessman betrayed by his partner lost a lot of money.

The executive betrayed by her accountant committed the famous tax fraud.

The policeman captured by the kidnapper divulged a lot of information.

The witch captured by the peasant cackled gleefully.

The suspect captured by the investigator looked terrified.

The prisoners captured by the colonel escaped at night.

The dog chased by the boy played with the ball.

The monkey chased by a hatter stole his hats.

The thief chased by the officer stopped suddenly.

The dentist chased by her son threatened playfully to extract his teeth.

The captain congratulated by the team jumped excitedly.

The actress congratulated by the director talked to the media happily.

The governor congratulated by the president made a public statement.

The valedictorian congratulated by her family beamed joyfully.

The widow consoled by her family felt better eventually.

The teenager consoled by the friend went back home sullenly.

The woman consoled by her daughter smiled wistfully.

The athlete consoled by his coach promised to put in more effort.

The boy described by the lady glowed joyfully.

The secretary described by the visitor waited patiently.

The chemist described by the writer was interested in philosophy.

The physicist described by the journalist published many seminal articles.

The gymnast encouraged by the coach participated in the contest.

The musician encouraged by the crowd performed a breathtaking piece.

The teacher encouraged by her students brought some donuts.

The hairdresser encouraged by her customer tried a new hairstyle.

The defendant examined by the lawyer was unreliable.

The engineer examined by a licensor passed the test.

The industrialist examined by the auditor hid the money.

The patient examined by the hygienist looked uncomfortable.

The thief identified by the victim ran away in fear.

The warden identified by the prisoner cracked the walnuts enthusiastically.

The student identified by the researcher proposed an interesting project.

The electrician identified by the company was competent.

The landlord loved by his tenants was considerate.

The comedian loved by the audience impersonated a famous celebrity.

The kid loved by her parents slept peacefully.

The janitor loved by the employees was cheerful.

The cashier paid by the customer muttered something inaudibly.

The organization paid by the government expanded its mission.

The broker paid by the company signed the contract.

The entrepreneur paid by the industrialist founded a successful company.

The man recognized by the spy took off down the street.

The actor recognized by the reporter waved enthusiastically.

The poet recognized by the writer smiled.

The driver recognized by the passenger grinned cheerfully.

The priest recorded by the mob preached non-violence.

The management recorded by the employee talked about privacy.

The dancer recorded by his trainer practiced energetically.

The influencer recorded by her fans started trending rapidly.

The girl scratched by the cat enjoyed the sunny afternoon.

The toddler scratched by the table exclaimed in surprise.

The dog scratched by a stranger barked loudly.

The diamond scratched by the machine remained flawless.

The applicant selected by the company gained more information.

The professor selected by the university made plans to move.

The chef selected by the restaurant advertised his signature dish.

The developer selected by the team solved the problem efficiently.

The culprit sketched by the investigator escaped.

The apprentice sketched by the artist rolled up the canvas.

The painter sketched by her student looked happy.

The robot sketched by the illustrator gained popularity rapidly.

The soldier studied by the general fought valiantly.

The writer studied by the scholar wrote a radical book.

The apprentice studied by the blacksmith forged a shield.

The scholars studied by the psychologist were attentive.

The officer envied by the assistant quit the job.

The principal envied by the administration complained incessantly.

The man envied by his sister bought a new car.

The plumber envied by the carpenter was well-known.

The speaker introduced by the chairman charmed everyone.

The producer introduced by the association won many awards.

The elf introduced by the magician danced happily.

therapist introduced by the doctor listened carefully.

The protesters challenged by the authority marched on the streets.

The scientist challenged by the journalist demonstrated the invention.

The dragon challenged by the wizard roared loudly.

The programmer challenged by the supervisor accepted her mistake.

References

Frazier, Lyn (2013). "Syntax in sentence processing". In: *Sentence processing* 446.

McRae, Ken and Kazunaga Matsuki (2013). "Constraint-based models of sentence processing". In: *Sentence processing* 519, pp. 51–77.

Levy, Roger (2013). "Memory and surprisal in human sentence processing". In: *Sentence processing* 519, pp. 78–115.

Spivey, Michael J., Sarah E. Anderson, and Thomas A. Farmer (2013). "Memory and surprisal in human sentence processing". In: *Sentence processing* 519, pp. 115–136.

Prasad, Grusha and Tal Linzen (2021). "Rapid syntactic adaptation in self-paced reading: detectable, but requires many participants." In: *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Chomsky, Noam et al. (1957). "Syntactic Structures". In: *The Hague: Mouton*.

Ross, John Robert (1967). "Constraints on variables in syntax." In.

Harwood, William (2018). "Reduced Relatives and Extended Phases: A Phase-Based Analysis of the Inflectional Restrictions on English Reduced Relative Clauses". In: *Studia Linguistica* 72.2, pp. 428–471.

Branigan, Holly P, Martin J Pickering, Simon P Liversedge, Andrew J Stewart, and Thomas P Urbach (1995). "Syntactic priming: Investigating the mental representation of language". In: *Journal of Psycholinguistic Research* 24.6, pp. 489–506.

Branigan, Holly P and Martin J Pickering (2017). "An experimental approach to linguistic representation". In: *Behavioral and Brain Sciences* 40.

Smith, Carlota S (1961). "A class of complex modifiers in English". In: *Language* 37.3, pp. 342–365.

Tooley, Kristen M, Martin J Pickering, and Matthew J Traxler (2019). "Lexically-mediated syntactic priming effects in comprehension: Sources of facilitation". In: *Quarterly Journal of Experimental Psychology* 72.9, pp. 2176–2196.

Fine, Alex B. and T. Florian Jaeger (2016). "The role of verb repetition in cumulative structural priming in comprehension". In: *Journal of Experimental Psychology:*

Learning, Memory, and Cognition 42.9, pp. 1362–1376. URL: http://dx.doi.org/10.1037/xlm0000236.

Kayne, Richard S (1994). *The antisymmetry of syntax*. Vol. 25. MIT press.

Bhatt, Rajesh (1999). "Covert modality in non-finite contexts: University of Pennsylvania dissertation". In.

Steedman, Mark (1996). *Surface Structure and Interpretation*. MIT Press.

Stabler, Edward P (2013). "Two models of minimalist, incremental syntactic analysis". In: *Topics in cognitive science* 5.3, pp. 611–633.

Baumann, Peter (2021). "Incremental Structure Building and Islands". PhD thesis. Northwestern University.

Stolcke, Andreas (1995). "An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities". In: *Computational Linguistics* 21.2, pp. 165–201. URL: https://aclanthology.org/J95-2002.

Hale, John (2001). "A Probabilistic Earley Parser As a Psycholinguistic Model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Pittsburgh, Pennsylvania: Association for Computational Linguistics, pp. 1–8. DOI: 10.3115/1073336.1073357. URL: https://doi.org/10.3115/1073336.1073357.

Gaston, Phoebe, Nick Huang, and Colin Phillips (2017). "The logic of syntactic priming and acceptability judgments". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000371.

Ziegler, Jayden, Jesse Snedeker, and Eva Wittenberg (2017). "Priming is swell, but its far from simple". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000607.

Kush, Dave and Brian Dillon (2021). "Sentence Processing and Syntactic Theory". In: *A Companion to Chomsky*, pp. 305–324.

Phillips, Colin, Phoebe Gaston, Nick Huang, and Hanna Muller (2021). "Theories all the way down: Remarks on "theoretical" and "experimental" linguistics". In: *The Cambridge handbook of experimental syntax*, pp. 587–616.

Anderson, John R, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin (2004). "An integrated theory of the mind." In: *Psychological review* 111.4, p. 1036.

Reitter, David, Frank Keller, and Johanna D Moore (2011). "A computational cognitive model of syntactic priming". In: *Cognitive science* 35.4, pp. 587–637.

Lewis, Richard L and Shravan Vasishth (2005). "An activation-based model of sentence processing as skilled memory retrieval". In: *Cognitive science*, pp. 375–419.

Lewis, Richard Lawrence (1993). *An architecturally-based theory of human sentence comprehension*. Carnegie Mellon University.

Lewis, Richard L (1998). "Reanalysis and limited repair parsing: Leaping off the garden path". In: *Reanalysis in sentence processing*. Springer, pp. 247–285.

Roark, Brian (2001). "Probabilistic top-down parsing and language modeling". In: *Computational linguistics* 27.2, pp. 249–276.

Ambati, Bharat Ram, Tejaswini Deoskar, Mark Johnson, and Mark Steedman (2015). "An incremental algorithm for transition-based CCG parsing". In: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 53–63.

Yang, Kaiyu and Jia Deng (2020). "Strongly incremental constituency parsing with graph neural networks". In: *Advances in Neural Information Processing Systems* 33, pp. 21687–21698.

Chang, Franklin, Gary S Dell, and Kathryn Bock (2006). "Becoming syntactic." In: *Psychological review* 113.2, p. 234.

Malhotra, Gaurav (2009). "Dynamics of structural priming". PhD thesis. University of Edinburgh.

Snider, Neal (2008). "Similarity and structural priming". PhD thesis. Stanford University.

Ruiter, Jan P. de and Laura E. de Ruiter (2017). "Don't shoot the giant whose shoulders we are standing on". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000346.

Koring, Loes and Eric Reuland (2017). "What structural priming can and cannot reveal". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000462.

Martin, Anrdrea E., Falk Huetting, and Mante S. Nieuwland (2017). "Can structural priming answer the important questions about language?" In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000528.

Rees, Alice and Lewis Bott (2017). "Structural priming is a useful but imperfect technique for studying all linguistic representations, including those of pragmatics". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000553.

Ryskin, Rachel and Sarah Brown-Schmidt (2017). "The malleability of linguistic representations poses a challenge to the priming-based experimental approach". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000577.

Steedman, Mark (2001). *The syntactic process*. MIT press.

Chomsky, Noam (1995). *The minimalist program*. MIT press.

Phillips, Colin (2013). "Parser-grammar relations: We don't understand everything twice". In: *Language down the garden path: The cognitive and biological basis for linguistic structures*, pp. 294–315.

Hockenmaier, Julia and Mark Steedman (2007). "CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank". In: *Computational Linguistics* 33.3, pp. 355–396. DOI: 10.1162/coli.2007.33.3.355. URL: https://aclanthology.org/J07-3004.

Adger, David (2003). *Core syntax: A minimalist approach*. Vol. 20. Oxford University Press Oxford.

Vasishth, Shravan and Felix Engelmann (2021). *Sentence comprehension as a cognitive process: A computational approach*. Cambridge University Press.

Clark, Stephen (2021). "Something Old, Something New: Grammar-based CCG Parsing with Transformer Models". In: *arXiv preprint arXiv:2109.10044*.

Clark, Stephen and James R Curran (2007). "Wide-coverage efficient statistical parsing with CCG and log-linear models". In: *Computational Linguistics* 33.4, pp. 493–552.

Xu, Wenduan, Michael Auli, and Stephen Clark (2015). "CCG supertagging with a recurrent neural network". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 250–255.

Tian, Yuanhe, Yan Song, and Fei Xia (2020). "Supertagging Combinatory Categorial Grammar with Attentive Graph Convolutional Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6037–6044. DOI: 10.18653/v1/2020.emnlp-main.487. URL: https://aclanthology.org/2020.emnlp-main.487.

Hockenmaier, Julia and Mark Steedman (2002). "Generative models for statistical parsing with Combinatory Categorial Grammar". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 335–342.

Zhang, Yue and Stephen Clark (2011). "Shift-reduce CCG parsing". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 683–692.

Steedman, Mark (1991). "Type-raising and directionality in combinatory grammar". In: *Technical Reports (CIS)*, p. 390.

Roland, Douglas, Frederic Dick, and Jeffrey L Elman (2007). "Frequency of basic English grammatical structures: A corpus analysis". In: *Journal of memory and language* 57.3, pp. 348–379.

Jeffreys, H (1939). *Theory of Probability*. Oxford University Press.

Fine, Alex, Ting Qian, T Florian Jaeger, and Robert Jacobs (2010). "Syntactic adaptation in language comprehension". In: *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pp. 18–26.

Delaney-Busch, Nathaniel, Emily Morgan, Ellen Lau, and Gina R Kuperberg (2019). "Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming". In: *Cognition* 187, pp. 10–20.

Schad, Daniel J, Bruno Nicenboim, Paul-Christian Bürkner, Michael Betancourt, and Shravan Vasishth (2022). "Workflow techniques for the robust use of bayes factors." In: *Psychological Methods*.

Mahowald, Kyle, Ariel James, Richard Futrell, and Edward Gibson (2017). "Structural priming is most useful when the conclusions are statistically robust". In: *Behavioral and Brain Sciences* 40. DOI: 10.1017/S0140525X17000504.

Makowski, Dominique, Mattan S. Ben-Shachar, and Daniel Lüdecke (2019a). "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework." In: *Journal of Open Source Software* 4.40, p. 1541. DOI: 10.21105/joss.01541. URL: https://joss.theoj.org/papers/10.21105/joss.01541.

Paape, Dario and Shravan Vasishth (2022). "Estimating the true cost of garden-pathing: A computational model of latent cognitive processes". In.

Makowski, Dominique, Mattan S. Ben-Shachar, and Daniel Lüdecke (2019b). "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework." In: *Journal of Open Source Software* 4.40, p. 1541. DOI: 10.21105/joss.01541. URL: https://joss.theoj.org/papers/10.21105/joss.01541.

Mahowald, Kyle, Ariel James, Richard Futrell, and Edward Gibson (2016). "A meta-analysis of syntactic priming in language production". In: *Journal of Memory and Language* 91, pp. 5–27.

Boston, Marisa Ferrara, John T Hale, Shravan Vasishth, and Reinhold Kliegl (2011). "Parallel processing and sentence comprehension difficulty". In: *Language and Cognitive Processes* 26.3, pp. 301–349.

Chomsky, Noam (1965). "Aspects ofthe theory ofsyntax". In: *Cambridge, MA: MIT-Press* 1977, pp. 71–132.

Ambridge, Ben (2017). "Horses for courses: When acceptability judgments are more suitable than structural priming (and vice versa)". In: *Behavioral and Brain Sciences* 40.

Adger, David (2017). "The limitations of structural priming are not the limits of linguistic theory". In: *Behavioral and Brain Sciences* 40.

Franck, Julie, Glenda Lassi, Ulrich H Frauenfelder, and Luigi Rizzi (2006). "Agreement and movement: A syntactic analysis of attraction". In: *Cognition* 101.1, pp. 173–216.

Kobele, Gregory M, Sabrina Gerth, and John Hale (2013). "Memory resource allocation in top-down minimalist parsing". In: *Formal grammar*. Springer, pp. 32–51.

Graf, Thomas, James Monette, and Chong Zhang (2017). "Relative clauses as a benchmark for Minimalist parsing". In: *Journal of Language Modelling* 5.1, pp. 57–106.

De Santo, Aniello (2021). "A Minimalist Approach to Facilitatory Effects in Stacked Relative Clauses". In: *Proceedings of the Society for Computation in Linguistics* 4.1, pp. 1–17.

Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen (2022). "SPR mega-benchmark shows surprisal tracks construction- but not item-level difficulty". In: *The 35th Annual Conference on Human Sentence Processing*.

Quine, Willard Van Orman (1960). *Word and object*. MIT press.

Chapter 3

Are the structures that the system of rules builds in temporarily ambiguous sentences impacted by context-specific probabilities?

This chapter was previously published as:

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

3.1 Introduction

Humans' ability to extract statistical regularities from their environment plays an important role in language acquisition and processing (Mitchell et al., 1995; Romberg and Saffran, 2010). In sentence comprehension, in particular, predictable syntactic structures are easier to process than unpredictable ones (MacDonald, Pearlmutter, and

Seidenberg, 1994; Trueswell, 1996). Under a rational account of sentence comprehension, we would expect these predictability effects to be driven by context-specific statistical regularities (Anderson, 1990): since the distribution of syntactic structures can vary widely across environments and contexts, readers' expectations will only be an accurate reflection of the statistics of the current environment if they can rapidly calibrate their expectations to match those statistics (Fine et al., 2013).

In line with this hypothesis, (Wells et al., 2009) showed that participants who were exposed to sentences with relative clauses over several experimental sessions read new sentences with relative clauses faster than did participants who were exposed to sentences with other syntactic structures. Building on this finding, (Fine et al., 2013) tested whether readers can recalibrate their expectations over the course of a single experimental session, focusing on sentences such as (1):

(1) The experienced soldiers warned about the dangers **conducted the midnight** raid. (Reduced RC; ambiguous)

Sentence (1) is temporarily ambiguous between a main verb reading, where the soldiers warned someone about the danger, and a relative clause reading, where the soldiers were warned by someone about the danger. The sentence is eventually disambiguated in favor of the relative clause reading by *conducted*. This temporary ambiguity is absent from a minimally different sentence with an unreduced relative clause like (2); in this sentence, only the relative clause reading is possible:

(2) The experienced soldiers who were told about the dangers **conducted the midnight** raid. (Unreduced RC; unambiguous)

Across a range of studies, the words of the disambiguating region of (1), marked in boldface, have been shown to be read more slowly than the same words in a matched unambiguous sentence such as (2) (MacDonald, Pearlmutter, and Seidenberg, 1994; Trueswell, 1996; Liversedge, Paterson, and Clayes, 2002; Clifton Jr et al., 2003; Kemper, Crow, and Kemtes, 2004). We refer to this difference in reading times as the garden path effect.

Fine et al. (2013) interpreted the garden path effect as a consequence of more general word predictability effects (following Hale (2001)): when reading the ambiguous region of sentence (1), participants are likely to interpret the verb *warned* as the main verb of the sentence, since verbs like *warned* occur more frequently as matrix clause verbs than as verbs introducing a passive reduced relative clause as in (1). Given this bias towards a main verb reading, words which disambiguate the temporarily ambiguous sentence in favor of the relative clause reading are less expected than the same words when they occur in a sentence like (2), where only a relative clause reading is possible. Since, all else being equal, less predictable words are read more slowly than predictable ones (Ehrlich and Rayner, 1981; Smith and Levy, 2013), the greater frequency of main verb parses can explain the garden path effect.

Fine and colleagues hypothesized that if participants update their expectations to match the statistics of the environment, then, in an experimental context where participants were exposed to several sentences such as (1), with reduced RCs, words that disambiguate the sentence in favor of the relative clause reading would become more predictable over time; this, in turn, would result in a decrease in the garden path effect. We will refer to this hypothesis as the *syntactic adaptation* hypothesis. In line with this hypothesis, Fine et al. (2013) observed a decrease in the garden path effect

over the course of a self-paced reading experiment, in which readers press a key to reveal the next word in the sentence. A similar decrease has since been observed in other self-paced reading studies (Fine and Jaeger, 2016; Stack, James, and Watson, 2018).

While the decrease in garden path effect is consistent with the syntactic adaptation account, syntactic adaptation is not the only possible explanation for this finding. In all of the studies mentioned above, as the experiment progressed, reading times (RTs) decreased not only for temporarily ambiguous sentences, but also for sentences in all other conditions, regardless of the syntactic structure of the sentence (Fine et al., 2013; Fine and Jaeger, 2016; Stack, James, and Watson, 2018). We will refer to the decrease in RTs that is independent of any recalibration of syntactic expectations as *task adaptation*. In the following paragraphs, we explain how task adaptation could result in a decrease in garden path effect, even in the absence of syntactic adaptation.

We assume that task adaptation does not directly depend on the syntactic structure of the sentence, but could depend on the speed with which the sentence is read when encountered early in the experiment. If the rate of task adaptation—the speedup in milliseconds from one trial to the next—is greater for sentences that are read more slowly at the beginning of the experiment (to which we will refer as "difficult sentences" for convenience) than for sentences that are read more rapidly ("easy sentences"), then, over time, the difference in RTs between easy and difficult sentences will decrease, resulting in a decrease in the garden path effect (see Figure 3.1). Such variability in difficulty across sentences could arise from any number of of factors, including word frequency, plausibility, predictability, and syntactic disambiguation difficulty. We will refer to the class of task adaptation functions that have this property as *start-point*

[h]

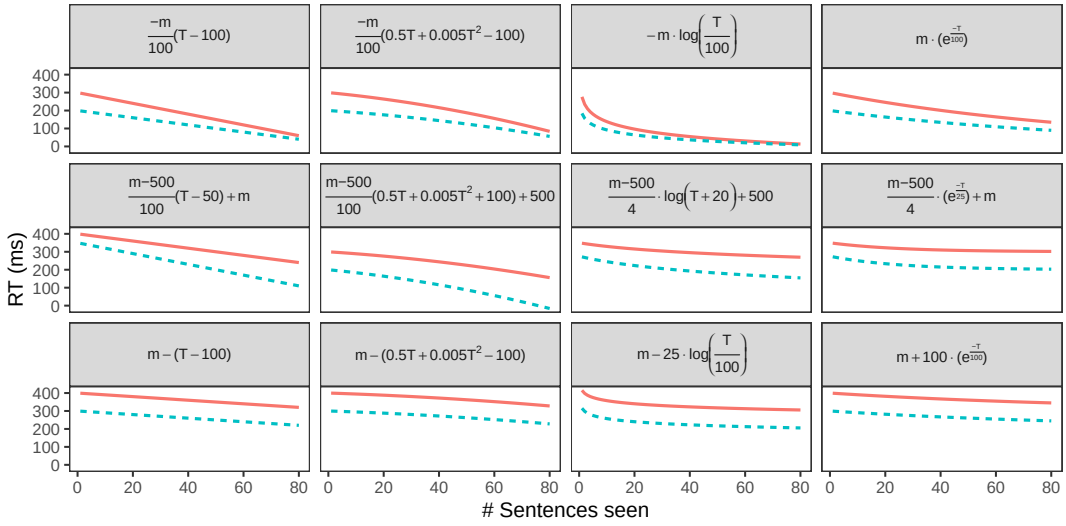


Figure 3.1: An illustration of some of the possible functions that could describe the decrease in reading time caused by task adaptation for two sentences (red solid and blue dashed) over the course of the experiment. At the beginning of the experiment (at trial 1), the sentence depicted by the red solid line is read more slowly than the sentence depicted by the blue dashed line. The top two rows depict functions that are sensitive to the initial reading times of the sentences (start-point dependent and diverging start-point dependent functions) and the bottom row depicts functions that are not sensitive to these initial reading times (start-point independent functions). The value of the parameter m is 300 for the red line and 200 for the blue one. The difference in RTs between the red solid and blue dashed line decreases only in the start-point dependent functions. These simple functions were chosen to illustrate the three classes of task-adaptation functions rather than for their psychological plausibility. While many of these functions are not psychologically plausible because they predict negative RTs after some trials, they can be modified to be more psychologically plausible (e.g., by enforcing a floor).

dependent task adaptation. If task adaptation is indeed start-point dependent, then even though the same task adaptation function applies to both reduced and unreduced RCs, the rate of decrease in RTs would be greater for reduced RCs than for unreduced RCs. If that is the case, it is possible that the decrease in garden path effect observed in previous studies was driven by task adaptation alone, or by a combination of task and syntactic adaptation.

There are at least two other possible types of task adaptation functions. First, the rate of task adaptation could be *lower* for difficult sentences than for easy ones (*diverging start-point dependent*). In this intuitively less likely case, the garden path effect would increase over time. Second, the rate of task adaptation could be identical for easy and difficult sentences (*start-point independent*). In this case, task adaptation would not cause the garden path effect to change over time. If task adaptation follows either of these patterns, the decrease in garden path effect observed by previous studies cannot be explained by task adaptation.

Since the form of the task adaptation function that characterizes self-paced reading studies is currently unknown, all of the three alternatives discussed above are possible. Therefore, we cannot know whether the decrease in garden path effect observed in previous studies was driven by start-point dependent task adaptation alone, by syntactic adaptation alone, or by a combination of the two. The goal of this paper is to adjudicate between these three possibilities. Before describing our approach, we briefly discuss previous attempts to do so.

The Fine et al. (2013) experiment mentioned above consisted of two blocks. In the first block, participants ($n = 80$) read either 16 filler sentences (Filler-exposed group), or 16 sentences with RCs, half of which had reduced RCs like (1), and the

other half unreduced RCs like (2) (RC-exposed group). Then, in the second block of the experiment, the garden path effect was measured in both groups by comparing the RTs for sentences with reduced RCs and with unreduced RCs (five each).[1] Fine et al. (2013) found that the garden path effect in the RC-exposed group decreased between the first block and the second. In the second block, the garden path effect was smaller in the RC-exposed group than the Filler-exposed group, although this interaction was only marginally significant ($\beta = -5$, $t = -1.7$, $p = 0.08$). Fine and colleagues argued that the decrease in garden path effect they observed was a result of syntactic adaptation: if it had been caused by task adaptation alone, the garden path effect would not differ across the two groups, both of which were exposed to the same number of sentences.

In a later experiment that used the same design as (Fine et al., 2013) but considerably more participants and items (423 participants, 32 sentences in Block 1 and 20 sentences in Block 2), (Stack, James, and Watson, 2018) replicated the decrease in the garden path effect observed by (Fine et al., 2013) for the RC-exposed group of participants, but failed to replicate the crucial interaction: the garden path effects in Block 2 did not differ significantly between the RC-exposed and Filler-exposed participants ($\beta = 1.25$, $t = 1.05$, $p > 0.05$).[2] Based on these results, Stack and colleagues argued that the observed decrease in garden path effect was likely driven by task adaptation and not by syntactic adaptation. In a response to Stack and colleagues, Jaeger, Bushong, and Burchill (2019) challenged these conclusions. Based

[1] Fine et al. (2013) also included a third block with sentences that were disambiguated in favor of the main verb reading, e.g., *The experienced soldiers warned about the dangers before the midnight raid.* We briefly discuss this manipulation in the General Discussion.

[2] The difference in signs is an artifact of how the predictors were coded in the two studies. In both the studies the garden path effect for the RC-exposed group was smaller than that for the Filler-exposed group.

on a reanalysis of the data from (Stack, James, and Watson, 2018) and computational simulations, Jaeger and colleagues argued that Stack, James, and Watson's experiment, far from being a failure to replicate their earlier work, in fact provides evidence for syntactic adaptation.

The present paper aims to clarify the empirical picture regarding syntactic adaptation in self-paced reading. We report on two experiments designed to investigate which of the factors described earlier can drive the decrease in garden path effect observed in self-paced reading experiments: will we observe syntactic adaptation only, task adaptation only, or a combination of the two? Instead of Fine et al. (2013), our design is based on the second experiment of (Fine and Jaeger, 2016) (henceforth referred to as FJ16); this experiment includes more items and has a simpler design than the earlier study by the same authors.[3] Across three similar experiments, FJ16 presented their participants with 20 sentences with reduced relative clauses (like (3a)) and 20 with unreduced relative clauses (like (3b)); as in (Fine et al., 2013), they found a decrease in the garden path effect over the course of the experiment.

(3) a. The evil genie served the golden figs went into a trance.

b. The evil genie who was served the golden figs went into a trance.

Experiment 1 of the present paper is a replication of FJ16. This replication had two goals: first, to ensure that the decrease in garden path effect can be replicated with FJ16's simpler design (to our knowledge, ours is the first attempt to replicate FJ16); and second, to investigate whether task adaptation is start-point dependent and,

[3]Specifically, FJ16 did not include the manipulation with sentences that were disambiguated in favor of the main verb reading, e.g., *The experienced soldiers warned about the dangers before the midnight raid.*

as such, can on its own lead to a decrease in garden path effect. This experiment successfully replicated the results of FJ16 in both direction and magnitude: as in FJ16, the garden path effect in our Experiment 1 decreased by approximately 1% with every additional reduced relative clause sentence encountered by the participant. We also found evidence that task adaptation is start-point dependent—the rate of task adaptation was greater for sentences that were initially read more slowly than for sentences that were initially read more rapidly. These results suggest that the observed decrease in garden path effect does not necessarily reflect syntactic adaptation: in principle, the decrease could have been driven entirely by start-point dependent task adaptation.

Next, Experiment 2 investigates whether syntactic adaptation results in a decrease in garden path effect over and above the decrease caused by start-point dependent task adaptation. Following a similar logic as in (Fine et al., 2013) and (Stack, James, and Watson, 2018), we used a between-group blocked design to compare the garden path effect between participants exposed to RRC sentences (RRC-exposed group) and those exposed to filler sentences (Filler-exposed group). As discussed earlier, if syntactic adaptation results in a decrease in garden path effect over and above task adaptation, we expect the garden path effect following exposure to be smaller in the RRC-exposed group than in the Filler-exposed group.

To test this prediction, we first ran a preliminary experiment, Experiment 2a, in which we measured the magnitude of the garden path effect in a Filler-exposed group. We then used this estimate to predict the magnitude of garden path effect that we are likely to observe for the RRC-exposed group. Based on this prediction, we ran a power analysis to estimate the number of participants required to detect between-group

difference in the garden path effect. This power analysis indicated that it would be possible to detect such an effect with adequate power with 800 participants. Next, in Experiment 2b, we collected data for both groups, with a sample size based on our power analysis, and found evidence for syntactic adaptation over and above task adaptation.

Finally, based on our data from Experiment 2b, we ran power analyses to estimate the number of participants required for future experiments investigating the effects of syntactic adaptation using similar between-group designs. These simulations suggested that self-paced reading experiments with a blocked between-group design identical to ours will require around 800 participants to detect the basic syntactic adaptation effect with adequate power; experiments aimed at detecting modulations of this basic effect—e.g., determining whether the magnitude of syntactic adaptation varies across RC types—could be underpowered even with 1200 participants. We conclude that while syntactic adaptation can be detected using self-paced reading (contra Stack, James, and Watson (2018)), this paradigm might not be very effective for studying this phenomenon; this explains the mixed results found in previous studies.

3.2 Experiment 1: Does the garden path effect decrease over time? Can task adaptation account for the decrease?

3.2.1 Method

3.2.1.1 Participants

We recruited 80 participants via Prolific, a crowdsourcing platform. All participants specified on their profile that English was their first language. They were compensated at a rate of $6.51 per hour.

3.2.1.2 Materials

We used the same 40 critical items and 80 filler sentences as FJ16. Each of the critical items had a reduced form as in (3a) and an unreduced form as in (3b). To avoid the temporary syntactic ambiguity illustrated in (3a), the main verbs in all filler sentences were verbs like *woke*, which can only be interpreted as a past tense verb (the past participle in this case would be *woken*), rather than verbs like *served*, which are ambiguous between the two forms.

We generated four pseudorandom orders and, for each of the four orders, two lists counterbalanced for sentence type (i.e. if list 1 had the unreduced version of sentence A and the reduced version of sentence B, list 2 would include the reduced version of sentence A and the unreduced version of sentence B). We then generated a reversed version of each of these eight lists, for a total of 16 lists. Each participant was assigned to one of these 16 lists. To ensure that stimuli from the three conditions—RRC sentences, URC sentences and filler sentences—were evenly distributed throughout the experiment, we generated the pseudorandom orders in five blocks, where each

block contained four RRCs, four URCs, and 16 filler sentences. Every two critical items were separated by at least one filler, and at most two critical items of the same condition were allowed to follow each other (across filler items).

3.2.1.3 Procedure

The experiment was hosted on the IbexFarm website (Drummond, 2016). The procedure was standard for self-paced reading experiments. At the beginning of every trial, each of the words of the sentence was replaced by a dash whose length was roughly equivalent to the length of the word. When the participant pressed the space bar, the dash was replaced by the next word in the sentence and the previous word disappeared. At the end of the sentence, the participant was presented with a comprehension question, and used the keys 'z' and 'm' to respond 'yes' and 'no' respectively. We used the same comprehension questions as FJ16. The correct answer was 'yes' half of the time. Before the experiment started, participants were asked to fill out a brief demographic survey, and were given three practice trials.[4]

3.2.2 Results

3.2.2.1 Data filtering and exclusion

Although we indicated that only workers whose first language is English should participate in the experiment, four participants reported that English was not their first language. We excluded these participants from our analyses. We further excluded three participants whose comprehension question accuracy on filler sentences was lower than 80%; we excluded from this calculation two fillers whose mean accuracy was two

[4]All the experiments described in this paper were approved by The Johns Hopkins University Homewood Institutional Review Board.

standard deviations lower than the mean accuracy across fillers. Since a majority of the comprehension questions did not directly test whether participants correctly parsed RRC sentences, we did not exclude trials in which participants responded incorrectly to the comprehension questions; our results were qualitatively similar when trials with incorrect answers were excluded.[5] Following the data exclusion criteria used by FJ16, all observations (words) with RTs lower than 100 ms or greater than 2000 ms were excluded. This lead to the exclusion of 0.47% of the observations from the participants who were not excluded.

3.2.2.2 Analysis 1.1: A replication of FJ16's analysis.

FJ16 divided each sentence into five regions: subject (*the experienced waitress*), relativizer (*who was*: only URC sentences had this region), ambiguous region (*cooked the grilled chicken*), disambiguating region (*sent her food*) and final word (*back.*). They log-transformed the RTs; further, to control for word length, they fit a linear mixed-effects model predicting log-transformed RTs from word length, and performed all subsequent statistical analyses on the residuals of this model.

Since the garden path effect, which is the focus of interest in the current work, manifests in the disambiguating region, we restricted our analysis of residualized log RTs to this region. We fit a linear mixed-effects model that was nearly identical to the one specified by FJ16 (we modified the random effect structure slightly in order to allow the model to converge).[6] The model included the following predictors:

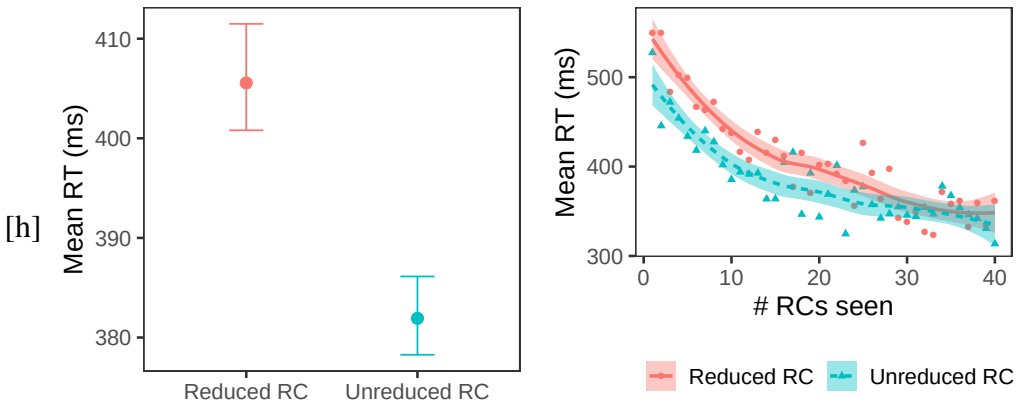
[5] We provide all details of analyses with the incorrect trials excluded in the following Open Science Framework (OSF) project: https://osf.io/57ckx/

[6] Fitting a model with the same random effect structure as in FJ16 yielded nearly identical $\hat{\beta}$ coefficients, but that model, unlike the model we report in this section, failed to converge. Further details can be found in the OSF project.

- Sentence type (referred to as Ambiguity in FJ16): A categorical variable coded as 1 for RRC sentences and -1 for URC sentences.
- Critical item number (Item order in FJ16): The number of critical items (reduced and unreduced) that the participant has seen so far.
- log(Stimulus number) (Stimulus order in FJ16): The natural log of the total number of sentences (critical items and filler sentences) that the participant has seen so far.
- Interaction between sentence type and critical item number.

Both critical item number and log stimulus order were centered around their mean. The model also included by-item and by-participant random intercepts, along with by-participant slopes for sentence type, critical item number and the interaction between sentence type and critical item number, as well as a by-item slope for sentence type. We estimated p values for the coefficients of this model using Satterthwaite's method, as implemented in the lmerTest package in R (Kuznetsova, Brockhoff, and Christensen, 2017).

The results of this analysis closely replicated FJ16. There was a significant garden path effect ($\hat{\beta} = 0.020$, $SE = 0.005$, $p \ll 0.01$; see Figure 3.2a). Length-corrected log RTs decreased significantly as a function of both log stimulus number ($\hat{\beta} = -0.083$, $SE = 0.008$, $p \ll 0.01$) and critical item number ($\hat{\beta} = -0.003$, $SE = 0.001$, $p = 0.02$). Crucially, the speedup over the course of the experiment was more pronounced for RRC sentences than for URC sentences ($\hat{\beta} = -0.001$, $SE = 0.0003$, $p < 0.01$; see Figure 3.2b). The coefficient of this interaction term was identical to that reported by FJ16 ($\hat{\beta} = -0.001$).



(a) **(b)**

Figure 3.2: Results of Experiment 1. (a) RTs in the disambiguating region for RRC sentences and URC sentences averaged over all participants and items. Error bars represent bootstrapped 95% confidence intervals. (b) RTs as a function of the number of critical items (both reduced and unreduced) seen by the participant, averaged across all participants and items. We fit the data points with a LOESS curve.

3.2.2.3 Analysis 1.2: Methods

This section reports an alternative analysis that addresses potential limitations of FJ16's analysis replicated in our Analysis 1.1. The first concern is that if word length is collinear with other predictors, then the residualization process used to correct for word length can bias the model's estimates and standard errors for the non-residualized predictors (York, 2012). Length correction is arguably unnecessary with the current design, which is within-item: since the critical region is identical across the URC and RRC versions of the same item, any effect of word length would be canceled out when we estimate the garden path effect. To address this potential issue, in Analysis 1.2 we used log-transformed RTs as the dependent variable instead of residualized length-corrected log transformed RTs used in Analysis 1.1.

A second concern regards the log transformation. The garden path effect is

typically calculated by summing or averaging RTs over the disambiguating region. But in Analysis 1.1 we averaged log-transformed RTs, which, when translated to the raw RT scale, is equivalent to *multiplying*, rather than summing, the RTs before dividing the log of the outcome by the number of words in the region. To avoid this counterintuitive arithmetic operation, in Analysis 1.2 we averaged the RTs in the disambiguating region *before* applying the log transformation.

Finally, Analysis 1.1 predicted log-transformed RTs as a linear function of log-transformed stimulus number; this is equivalent to assuming a linear relationship between RTs and stimulus number. Previous work outside of the sentence processing literature, however, suggests that RTs decrease exponentially, not linearly, as a function of practice (Heathcote, Brown, and Mewhort, 2000). In Analysis 1.2, we avoided log-transforming our stimulus number predictor; as a result, this analysis assumes a linear relationship between log-transformed RTs and stimulus number, or, equivalently, an exponential relationship between raw RTs and stimulus number, in line with prior work on the effect of practice.

In summary, the model we fit in Analysis 1.2 included the following predictors: stimulus number, ambiguity, critical item number, and the interaction between ambiguity and critical item number. We centered both stimulus number and critical item number by their mean and scaled them by their standard deviation. The random effect structure for this model included by-item and by-participant random intercepts, along with by-participant and by-item slopes for ambiguity, critical item number and the interaction between the two. We were unable to include by-item and by-participant random slopes for stimulus-number due to model convergence issues.

3.2.2.4 Analysis 1.2: Results

In this analysis, unlike in Analysis 1.1, the overall decrease in RTs across all conditions was only marginally significant ($\hat{\beta} = -0.158$, $SE = 0.091$, $p = 0.08$). Crucially, however, the magnitude of the garden path effect was greater than in Analysis 1.1, as was the magnitude of the decrease in the garden path effect (garden path effect: $\hat{\beta} = 0.024$, $SE = 0.005$, $p \ll 0.01$; decrease in garden path effect: $\hat{\beta} = -0.014$, $SE = 0.004$, $p \ll 0.01$). If anything, then, addressing our concerns with FJ16's analytical choices caused the effects of primary interest to be more pronounced than they were in Analysis 1.1.

3.2.3 Is task adaptation start-point dependent?

The decrease in RTs across all conditions as a function of stimulus number that was observed in analyses 1.1 and 1.2 suggests, in line with previous studies (Fine et al., 2013; Fine and Jaeger, 2016; Stack, James, and Watson, 2018), that participants adapt to the self-paced reading paradigm and read sentences more rapidly as the experiment progresses. However, these results do not directly speak to the question of whether task adaptation is start-point dependent or start-point independent[7]—i.e. whether or not the rate of task adaptation is greater for sentences that are read relatively slowly when presented early in the experiment ("difficult sentences") than for those that are read relatively rapidly when presented early in the experiment ("easy sentences"). As discussed earlier, if task-adaptation were indeed start-point dependent, we expect the difference in RTs between easy and difficult sentences to decrease over time, raising the possibility that the decrease in garden path effect observed in Experiment 1 was driven entirely by start-point dependent task adaptation. In this section we investigate

whether task adaptation is in fact start-point dependent.

We define the difficulty of a sentence x, which we denote $RT_{start}(x)$, as the time taken to read a word in sentence x, averaged across all the words in x and across all participants, when x was one of the first 24 sentences presented in the experiment (i.e. in the first block of the experiment).[8] Similarly, we define $RT_{end}(x)$ as the average RT on x when x was one of the last 24 sentences presented in the experiment (i.e. in the last block of the experiment). We then define $\Delta RT(x)$, the rate of task adaptation measured on x, as follows:

$$\Delta RT(x) = RT_{start}(x) - RT_{end}(x)$$

If task adaptation is start-point dependent, then for two sentences x and y where $RT_{start}(x) > RT_{start}(y)$ (i.e., x is more difficult than y), we would expect $\Delta RT(x) > \Delta RT(y)$.

To estimate ΔRT for all sentences, we first randomly split our participants into two halves. We used the first half of the participants (the *Difficulty Estimation Group*) to bin sentences according to their difficulty. Then, using the second half of the participants (*Task Adaptation Estimation Group*), we measured the rate of task adaptation by comparing the RTs at the start and end of the experiment for the sentences included in each bin. We used two sets of participants in this manner to avoid a circular analysis where the process of grouping sentences by their difficulty biases our estimates of task adaptation.

[7] Given the data, it is unlikely that task adaptation is characterized by diverging start-point dependent task-adaptation because these functions predict an *increase* in garden path effect over time, whereas we observed a decrease.

[8] Our definition of difficulty is empirical and is agnostic to *why* a particular sentence is difficult a priori. In future work, alternative definitions could categorize sentences based on factors such as word length or frequency, syntactic complexity, and so on.

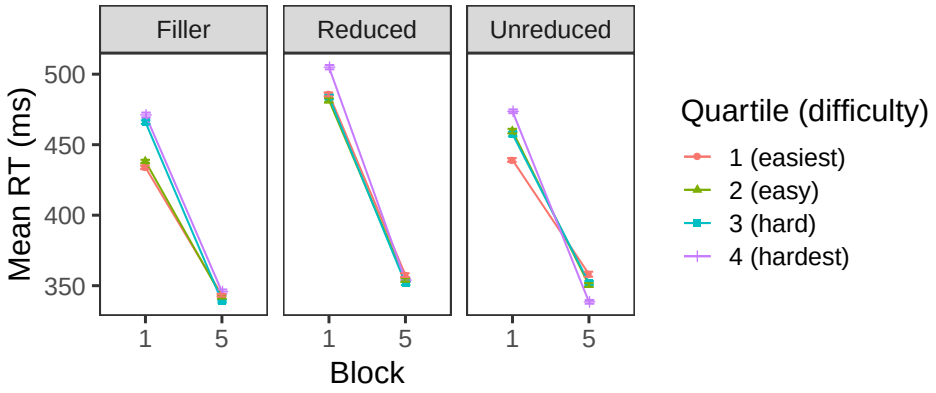


Figure 3.3: Task adaptation in Experiment 1. We plot RTs for participants in the Task Adaptation Estimation Group averaged across all words in the sentence for all sentences in Block 1 and Block 5. Sentences are binned into quartiles based on the RTs in Block 1 for participants in the Difficulty Estimation Group (binning was performed separately for each of the three classes of sentences). The estimates are averaged across 1000 random splits of participants. Error bars reflect two standard errors above and below the mean.

The analysis proceeded as follows. Using the RTs for the participants in the Difficulty Estimation Group, we computed RT_{start} for each filler sentence. Then, we binned these sentences into quartiles based on their RT_{start} values only (without taking into account their RT_{end}): for example, the first quartile consisted of the 25% of the sentences that were read most rapidly in Block 1 by the participants in the Difficulty Estimation Group, and the fourth quartile consisted of the 25% of sentences that were read most slowly in Block 1. We repeated this process separately for RRC, URC and filler sentences. Then, using the RTs from the other half of participants—the Task Adaptation Estimation Group—we computed the mean RT_{start} and RT_{end} for each quartile and for each of the three types of sentences by averaging the RTs for all words in all of the sentences included in that quartile. We repeated this process for 1000 random splits of participants, and averaged our RT_{start} and RT_{end} estimates across these random splits.

The results of this analysis indicate that in our data task adaptation was indeed start-point dependent (Figure 3.3): ΔRT was greater for sentences that were read more slowly when presented early in the experiment than for sentences that were read more rapidly. Difficulty was generally consistent across the Difficulty Estimation Group and the Task Adaptation Estimation Group. This pattern held for filler sentences as well as for RRC and URC sentences. Crucially, on average, ΔRT was greater for RRC sentences then URC sentences; this leads to a decrease in the difference in RTs between RRC sentences and URC sentences over the course of the experiment. In other words, at least some of the decrease in garden path effect over time observed in Experiment 1 can be accounted for by start-point dependent task adaptation.[9]

3.2.4 Discussion

Experiment 1 had two goals. The first was to replicate the decrease in garden path effect observed in previous studies. The second goal was to determine whether task adaptation is start-point dependent: if it is, then it could account at least in part for any decrease in garden path effect. We replicated in both direction and magnitude the decrease over time in garden path effect that was reported by FJ16; the coefficient of the interaction between sentence type and critical item number was -0.001 in both cases. This increases our confidence in the robustness of FJ16's empirical finding. At the same time, we also found that the decrease in RT measured for a particular sentence—whether it was an RRC, URC or filler sentence—depended on its "difficulty", or the time participants took on average to read that sentence when they encountered it early in the experiment. This suggests that at least a part of the observed

[9]We observed qualitatively similar results when we repeated the analysis with log transformed RTs. This analysis can be found on OSF.

Group	Exposure phase	Test phase
RRC-exposed	16 RRC, 16 Fillers	12 RRC, 12 URC, 24 Fillers
Filler-exposed	32 Fillers	12 RRC, 12 URC, 24 Fillers

Table 3.1: Design of Experiment 2. Experiment 2a only included a Filler-exposed group, whereas Experiment 2b included both groups.

decrease in garden path effect was driven by start-point dependent task-adaptation. In any study whose goal is to measure syntactic adaptation, then, it is essential to demonstrate that exposure to a certain syntactic structure results in a decrease in garden path effect *over and above* the decrease caused by task adaptation alone. The following section describes experiments motivated by this goal.

3.3 Overview of Experiments 2a and 2b

As discussed earlier, the syntactic adaptation account predicts that participants exposed to reduced relative clauses early in the experiment will be less surprised when reading these structures later on in the experiment, and will consequently display a reduced garden path effect compared to participants who are not exposed to sentences without such relative clauses early in the experiment. We test this prediction using a between-subject design with two phases, an exposure phase and a test phase (the division between the two phrases was not indicated to participants). In the exposure phase, participants in the RRC-exposed group read both RRC and filler sentences, whereas participants in the Filler-exposed group read only filler sentences. In the test phase, both groups of participants read RRC sentences, URC sentences, and filler sentences. This design is summarized in Table 3.1.

We ran two experiments using this design. In Experiment 2a, we collected data

from 81 participants, all of which were assigned to the Filler-exposed group. We used this smaller preliminary experiment to obtain an estimate of the garden path effect that arises in a setting where only task adaptation is possible. We then used the results of Experiment 2a as a basis for simulations whose goal was to predict the garden path effect for the RRC-exposed group, where both task adaptation and syntactic adaptation are at least in principle possible. Based on these estimates, we conducted power simulations whose goal was to estimate the number of participants required to reliably detect a between-group difference in the garden path effect; we then ran that number of participants in Experiment 2b.

3.4 Experiment 2a: What is the garden path effect for Filler-exposed participants?

3.4.1 Methods

3.4.1.1 Participants

We recruited 81 participants from Amazon's Mechanical Turk (one participant recruited unintentionally). This number was nearly identical to the number of participants recruited in FJ16 and in Experiment 1 (80). To limit the number of non-native speakers, participants were only recruited if the home address associated with their Amazon account was located in the United States. We based the compensation for our participants on a $8/hour rate (which was 75 cents above the US minimum wage at the time the experiment was run). Since the average duration of the experiment was approximately 15 minutes, participants received $2 for their time.

3.4.1.2 Materials

Our materials were based on those of FJ16, with two modifications. First, we added the word *the* to the beginning of four of FJ16's original stimuli, to ensure consistency across all items. Second, we replaced 27 of FJ16's original sentences with new ones. We did so because some of FJ16's sentences had verbs with a transitivity bias—that is, verbs that typically occur with a noun phrase (NP) complement—which caused them to be effectively disambiguated before the start of the disambiguating region (cf. Malone and Mauner (2018)). The following sentence from F16's materials, for example, is in practice disambiguated in favor of the relative clause reading at the prepositional phrase (*in the alley*), rather than at the second verb (*ran*), as intended:

(4) The calico cat licked in the alley ran into the street.

After the preposition phrase *in the alley* is encountered, a main verb reading can only be maintained under a heavy NP shift parse (e.g., *the cat licked in the alley the toy*). Since heavy NP shifts are relatively infrequent, the relative clause reading becomes highly probable even before the disambiguating region. This is likely to diminish the garden path effect in the disambiguating region in such sentences, and, consequently, diminish the extent to which they will cause syntactic adaptation—and thereby our power to detect a syntactic adaptation effect. We replaced these items with sentences that included optionally reflexive verbs (5a), ditransitive verbs (5b), or optionally transitive verbs without a strong transitivity bias (5c), where transitivity bias was determined based on estimates from (Roland and Jurafsky, 2002):

(5) a. The bearded man shaved two weeks ago liked his stylish new look.

b. The helpful librarian lent the frayed book took good care of it.

c. The ferocious lions attacked during the day were unable to escape the hunters.

After both of these modifications, all the sentences had seven words before the disambiguating region: three words in the subject NP, one verb and three words in the NP or prepositional phrase following the verb. We also created 64 filler sentences with similar properties to those we used in Experiment 1: they did not contain any relative clauses, and the main verbs' past participle differed from their past tense form.

3.4.1.3 Design

Experiment 2a consisted of an exposure phase and a test phase. In the exposure phase, participants read 32 filler sentences. In the test phase, they were presented with 12 RRC sentences, 12 URC sentences and 24 filler sentences (see Table 3.1). We generated four pseudo-random orders and two lists from each order, counter-balanced for ambiguity in the test phase, as in Experiment 1.

3.4.1.4 Procedure

The same procedure was used as in Experiment 1.

3.4.2 Results

3.4.2.1 Data filtering and exclusion

We used the same filtering and exclusion criteria as in Experiment 1. We excluded one participant who reported that English was not their first language. We additionally excluded eight participants whose mean accuracy on filler sentences was lower than

80%. Finally, we excluded all observations (words) with reading times lower than 100 ms or greater than 2000 ms, leading to the exclusion of 0.36% of all observations of the participants who were not excluded.

3.4.2.2 Estimating the garden path effect in the test phase

For every participant and trial, we averaged the RTs on the words in the disambiguating region. We then used a linear mixed-effects model to predict the log of these averaged RTs from sentence type (coded as 1 for RRC sentences and -1 for URC sentences). As discussed in Analysis 1.2, we did not include word length as a predictor because the critical region contained the same words across the RRC and URC version of a given item. We used the maximal random effects structure: by-participant and by-item random intercepts and a by-participant random slope for sentence type.

This model revealed a significant garden path effect: the disambiguating region was read significantly more slowly in RRC sentences than in URC sentences ($\hat{\beta} = 0.015$, $SE = 0.006$, $p = 0.02$).

3.4.3 Power analysis for Experiment 2b

Before conducting Experiment 2b, which follows the between-group design described above, we conducted simulations to estimate the number of participants required to obtain at least 80% power in this paradigm. We expect to observe a greater garden path effect when only task adaptation is possible (in the Filler-exposed group) than when both task adaptation and syntactic adaptation are possible (for the RRC-exposed group). To estimate power, we need a hypothesis as to the relative magnitude of the garden path effect size for each group, or the value of Ω in $GPE_{RRC} = \Omega \cdot GPE_{Filler}$,

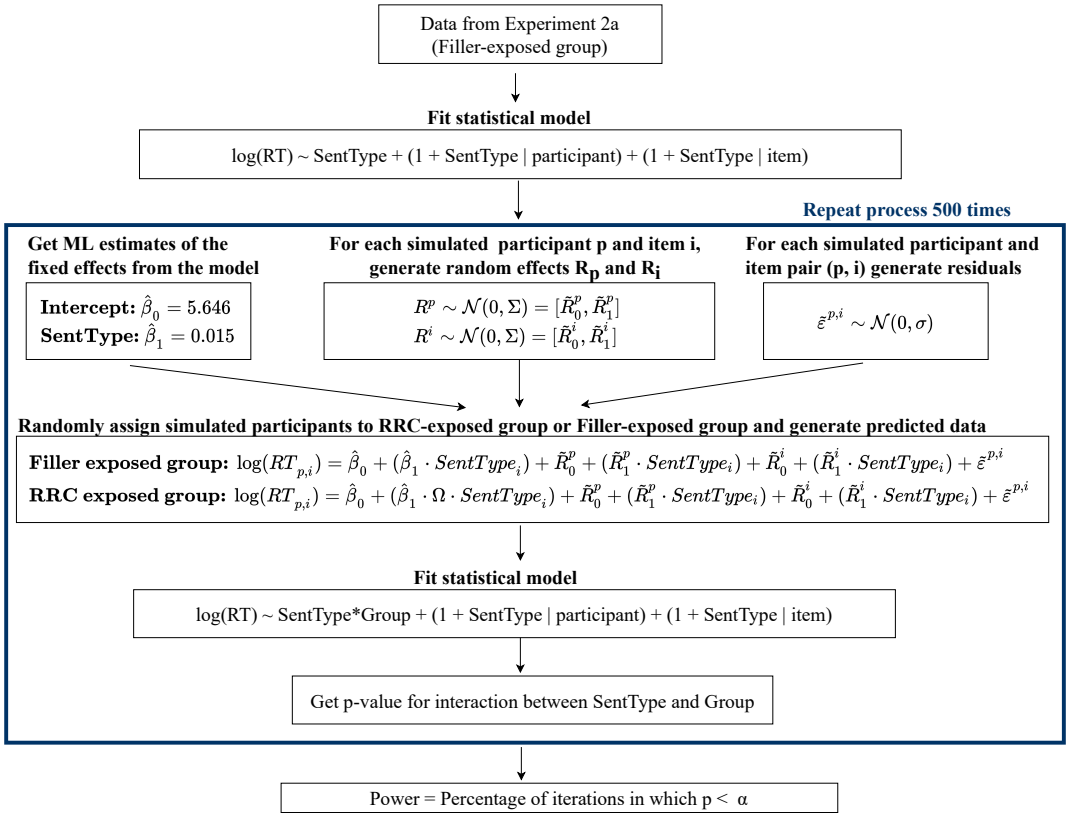


Figure 3.4: A schematic of how we calculated the power to detect a significant difference in the garden path effect between the RRC-exposed group and the Filler-exposed group. We use the LMER notation in R for Model_1 and Model_2. The fixed effects for Model_2, ($\hat{\beta}_0$ and $\hat{\beta}_1$), were estimated from Experiment 2a, and correspond to the coefficients of the intercept and sentence type respectively. The by-participant and by-item random intercepts ($\tilde{R}_0^p$, $\tilde{R}_0^i$) and random slopes ($\tilde{R}_1^p$, $\tilde{R}_1^i$), were sampled from the multivariate normal distribution $\mathcal{N}(0, \Sigma)$ where Σ corresponds to the covariance matrix of Model_1. The residual error for each observation ($\tilde{\varepsilon}^{p,i}$) was sampled from the normal distribution $\mathcal{N}(0, \sigma)$, where σ corresponds to the residual standard deviation of Model_1.

Ω value	# Participants	$p < 0.05$	$p < 0.01$	$p < 0.001$
0.10	200	0.45	0.21	0.05
	400	0.76	0.55	0.22
	800	0.97	0.89	0.68
0.18	200	0.38	0.16	0.04
	400	0.68	0.42	0.14
	800	0.94	0.82	0.54
0.25	200	0.31	0.13	0.04
	400	0.60	0.34	0.09
	800	0.89	0.73	0.44
0.50	200	0.17	0.05	0.01
	400	0.29	0.10	0.01
	800	0.59	0.34	0.09

Table 3.2: Power to detect a significant difference in the garden path effect between a Filler-exposed group and an RRC-exposed group if the garden path effect of the RRC-exposed group was 0.18 times that of the Filler-exposed group.

where GPE_{RRC} denotes the garden path effect for the RRC-exposed group, GPE_{Filler} the garden path effect for the Filler-exposed group, and $\Omega < 1$ is a constant proportion.

The simulations we report below are based on $\Omega = 0.18$; this value was derived from a simple Bayesian belief update model (Fine et al., 2010). After running Experiment 2b, we discovered an error in the calculation; however, post-hoc power calculations with other values of Ω revealed that the estimates for the number of required participants did not change substantially for values up to $\Omega = 0.25$ (see Table 3.2).

To estimate the power of our paradigm to detect a between-group difference in the garden path effect with n participants and the number of items included in the experiment, we sampled participants and items from the empirical random effect distribution estimated in Experiment 2a. We then randomly assigned half of the

simulated participants to the Filler-exposed group and the other half to the RRC-exposed group. For the Filler-exposed group, we generated predicted RTs for each trial by combining the fixed and random effects estimates from Experiment 2a with a sample from the same model's residual distribution. For the RRC-exposed group, we used a similar process but with one difference: we multiplied the coefficient of Sentence type (i.e., the garden path effect) by Ω.

With this simulated dataset in place, we then fit a linear mixed-effects model whose fixed effects included *Sentence type* (coded 1 for RRC sentences and -1 for URC sentences), *Group* (coded 1 for the RRC-exposed group and -1 for the Filler-exposed group), and the interaction between these two predictors. The random effects included intercepts for participants and items, along with a by-item and by-participant slope for Sentence Type. The random effect structure was not maximal because it was not possible to include a by-item slope for group: since Experiment 2a did not include RRC-exposed participants, we could not estimate the by-item variability in the difference between the two groups. Finally, we calculated the p value for the crucial interaction between Sentence Type and Group. For a diagram summarizing this procedure, see Figure 3.4.

We repeated the above process 500 times each for 200, 400 and 800 participants and for four different values of Ω: 0.10, 0.18, 0.25 and 0.50.[10] Table 3.2 summarizes the percentage of iterations in which the interaction between Sentence Type and

[10] A reviewer points out that 500 iterations for each combination of Ω and n are insufficient to obtain precise estimates—assuming a binomial distribution for the power estimates, with 500 iterations, it is not unlikely that our power estimates differed from the true value by up to 10 percentage points (i.e., if our estimate was 0.8, then it the true power likely lies between 0.9 and 0.7). The lack of precision does not change our conclusions, since even if the true power with 800 participants were 10 percentage points lower, the power would still be greater than 0.8; however, we recommend that in future work a larger number of iterations is used.

Group was significant for each of the datasets at different p value thresholds for rejecting the null hypothesis (α levels). Our power simulations indicate that for values of Ω up to 0.25 (i.e, if the garden path effect of the RRC-exposed group is predicted, under the syntactic adaptation hypothesis, to be a quarter of that of the Filler-exposed group) the power to detect a significant interaction was greater than 0.9 with 800 participants. One striking finding is that at $\alpha = 0.05$, the power to detect a significant interaction was much lower than 0.8 even with 200 participants—far more than typically participate in self-paced reading experiments.

3.5 Experiment 2b: Is the garden path effect for the Filler-exposed group greater than that for the RRC-exposed group?

3.5.1 Methods

3.5.1.1 Participants

We recruited participants on Amazon's Mechanical Turk using Microbatcher (Leonard, 2019). We planned to include in the experiment 800 participants, but ended up recruiting a slightly larger number (828). Only participants whose home address was located in the United States were recruited. Participants received $2 for their time.

3.5.1.2 Materials and Design

We used the same materials as in Experiment 2a. Filler-exposed participants were randomly assigned to one of the eight lists generated from the four pseudo-random orders used in Experiment 2a. We created eight additional lists for the RRC-exposed group by replacing 16 of the fillers from the exposure phase with RRC sentences.

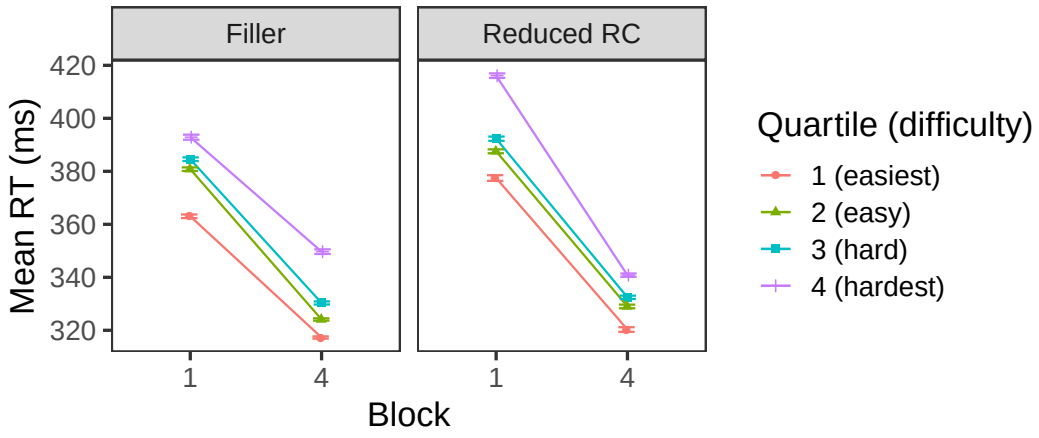


Figure 3.5: RTs for participants in the Task Adaptation Estimation Group averaged across all words in the sentence for all sentences in Block 1 and Block 5. Sentences (both critical items and filler sentences) are grouped into quartiles based on the RTs in Block 1 for participants in the Difficulty Estimation Group. Estimates are averaged across 1000 random splits of participants, and error bars reflect two standard errors above and below the mean.

RRC-exposed participants were randomly assigned to one of the latter eight lists.

3.5.1.3 Procedure

The procedure was identical to Experiments 1 and 2a.

3.5.2 Results

3.5.2.1 Data filtering and exclusion

We used the same data filtering and exclusion criteria as in Experiment 2a. This led to the exclusion of 11 participants who reported that English was not their first language and 175 participants whose accuracy on filler sentences was lower than 80%. The high proportion of participants with low filler accuracy in comparison to Experiment 2a cannot be attributed to question difficulty: in both experiments, Filler-exposed participants were presented with the same fillers, yet the proportion of

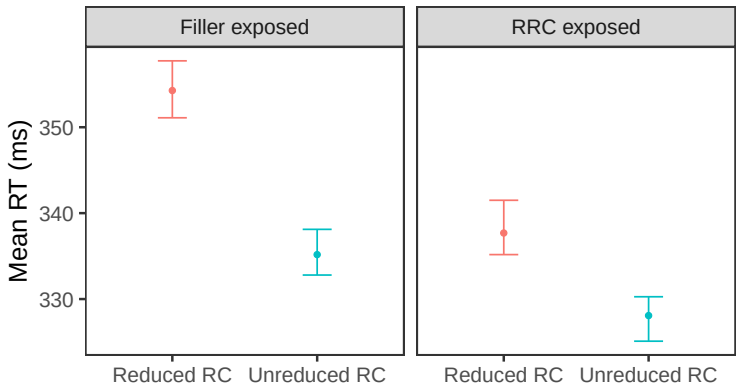


Figure 3.6: Garden path effect in the test phase for the Filler-exposed group and RRC-exposed group. Error bars reflect bootstrapped 95% confidence intervals.

participants with low filler accuracy differed drastically between the two experiments (10% in Experiment 2a and 21% in Experiment 2b). Additionally, even though the RRC-exposed group was presented with just a subset of the fillers presented to Filler-exposed group, the number of participants whose accuracy was low did not differ between the groups (87 in the Filler-exposed group and 88 in the RRC-exposed group), further suggesting that the difference in accuracy was not driven by the presence or absence of specific items. It is possible that the larger sample size of Experiment 2b led to the recruitment of less attentive participants or even bots.

As in the previous experiments, we also excluded observations (words) with RTs less than 100 ms or greater than 2000 ms. This led to the exclusion of 0.48% of all observations for the remaining 642 participants.

3.5.2.2 Is the rate of task adaptation higher for more difficult items?

We used the same method to diagnose start-point dependent task adaptation as in Experiment 1. We sampled half of the participants, and we divided both RRC and filler sentences into quartiles based on their RTs in this group of participants prior to task

adaptation (that is, early in the experiment). Then, using the remaining participants, we estimated the rate of task adaptation for each quartile by comparing the mean RTs, averaged across all sentences in the quartile, before and after task adaptation. We repeated this process for 1000 random splits of participants. As in Experiment 1, in almost all quartiles and types of sentences, sentences that were read more slowly when presented early in the experiment showed a greater task adaptation effect (ΔRT) than sentences that were read more rapidly early in the experiment. This supports the hypothesis that task adaptation is start-point dependent (see Figure 3.5).[11] As discussed earlier, we expect the rate of start-point dependent task adaptation to be similar across RRC-exposed and Filler-exposed participants. As such, a difference between groups in garden path effect in the test phase can only be attributed to syntactic adaptation.

3.5.2.3 Is there evidence for syntactic adaptation over and above task adaptation?

As in Experiment 2a, we averaged the RTs in the disambiguating region and log-transformed these averaged RTs. We then fit a linear mixed-effects model with the predictors we used in our power simulations. The fixed effects included Sentence Type, Group and the interaction between the two; and the random effects included random intercepts for participants and items, along with a by-participant slope for sentence type and by-item slope for sentence type, group and the interaction between the two.

[11] The only exception were the filler sentences that were read the most slowly (i.e., in the fourth quartile). For these sentences, ΔRT was smaller than for other filler sentences that were read more rapidly. We find qualitatively similar results when we repeat this analysis with log transformed RTs, with the exception of the RRC sentences that were read most rapidly, where ΔRT was larger than for other RRC sentences that were read more slowly. This analysis can be found on OSF.

The model revealed a significant garden path main effect: the words in the disambiguating region were read more slowly in RRC sentences than in URC sentences ($\hat{\beta} = 0.016$, $SE = 0.002$, $p \ll 0.001$). There was also a main effect of group: Filler-exposed participants read sentences significantly more slowly on average than RRC-exposed participants ($\hat{\beta} = 0.038$, $SE = 0.010$, $p < 0.001$). We briefly discuss this effect, which is not predicted by the syntactic adaptation hypothesis, in the discussion section. Finally, the crucial interaction was significant: the garden path effect was greater for the Filler-exposed group than for the RRC-exposed group ($\hat{\beta} = 0.006$, $SE = 0.002$, $p = 0.001$), providing evidence for syntactic adaptation over and above task adaptation (see Figure 3.6).

As was pointed out by a reviewer, by fitting a linear mixed-effects model to log transformed RTs, we made the (standard) assumption that RTs are lognormally distributed, and therefore assumed that the lowest possible RT was 0 ms. This assumption is physiologically implausible: RTs are constrained by factors such as the speed of muscle movements and cannot in practice be as low as 0 or 1 ms. To address this issue, we reanalyzed the data using Bayesian mixed-effects models based on the assumption that RTs follow a shifted log normal distribution (Rouder, 2005), a generalized form of the lognormal distribution with a shift parameter which determines the lowest possible RT value that the model can predict (i.e. the floor). The fixed effect and random effect structure of the shifted model was identical to the unshifted model described above. We allowed the shift parameter of the lognormal distribution to vary across participants. We used weakly informative priors, as recommended by Schad, Betancourt, and Vasishth (2019). These priors expressed the assumptions that RTs are very likely to lie between 100 to 2000 ms, and that the difference in RTs

between RRC and URC sentences was likely to lie between -100 and 100 ms, as was the difference in garden path effect between the RRC-exposed and Filler-exposed groups.[12]

The shifted model revealed qualitatively similar effects to the unshifted model, although all of the fixed effects were larger and there was more uncertainty about the estimates: a garden path main effect ($\hat{\beta} = 0.033$, $SE = 0.006$), a main effect of group ($\hat{\beta} = 0.062$, $SE = 0.018$), and an interaction between group and garden path effect ($\hat{\beta} = 0.009$, $SE = 0.004$).

3.5.3 Discussion

As in Experiment 1, we found that the effect of task adaptation was start-point dependent—the rate of decrease in RTs was higher in sentences that were read slowly when presented early in the experiment than sentences that were read rapidly. This supports the hypothesis that task adaptation causes a decrease in the garden path effect over time. At the same time, we also found evidence for a decrease in garden path effect over and above the decrease caused by task adaptation—the garden path effect was greater in participants who were only exposed to filler sentences in the exposure phase than in those who were exposed to 16 RRC sentences. This lends support to the syntactic adaptation hypothesis. However, as we discuss below, this effect is relatively small; this fact, in conjunction with design decisions that could have led to reduced power, may explain the recent failure of (Stack, James, and Watson, 2018) to observe a syntactic adaptation effect.

We also found that Filler-exposed participants read sentences significantly more

[12] Further details about the priors can be found on OSF.

slowly on average than participants in the RRC-exposed groups (see Figure 3.6). A similar main effect of group, which is not predicted by the syntactic adaptation account, was observed by both (Fine et al., 2013) and (Stack, James, and Watson, 2018). One possible explanation for this finding is that extensive exposure to syntactically simple filler sentences, followed by a sudden transition to syntactically challenging RRC sentences in the test phase, causes Filler-exposed participants to slow down and read all test sentences more carefully. Future work can test this hypothesis by determining whether this pattern persists when the Filler-exposed group is exposed to sentences that include temporary syntactic ambiguities other than that used to measure the garden path effect, for example the direct object / sentential complement (NP/S) ambiguity if the target ambiguity is main verb / reduced relative as in the present study.

3.5.3.1 Exploratory analyses

We now turn to exploratory analyses that further investigate the viability of self-paced reading as a paradigm for studying syntactic adaptation. We estimate the number of participants required for future experiments using this paradigm and compare the magnitude of task adaptation and syntactic adaptation.

3.5.3.2 How many participants should be recruited for future experiments with the same design?

This section reports the results of simulations whose goal was to estimate the power to detect a between-group difference in the garden path effect in future experiments with the same design as Experiment 2b. This approach was similar to the power analysis we conducted using the data from Experiment 2a, with two crucial differences. First, in Experiment 2a, we fit a linear mixed-effects model and calculated the power based on

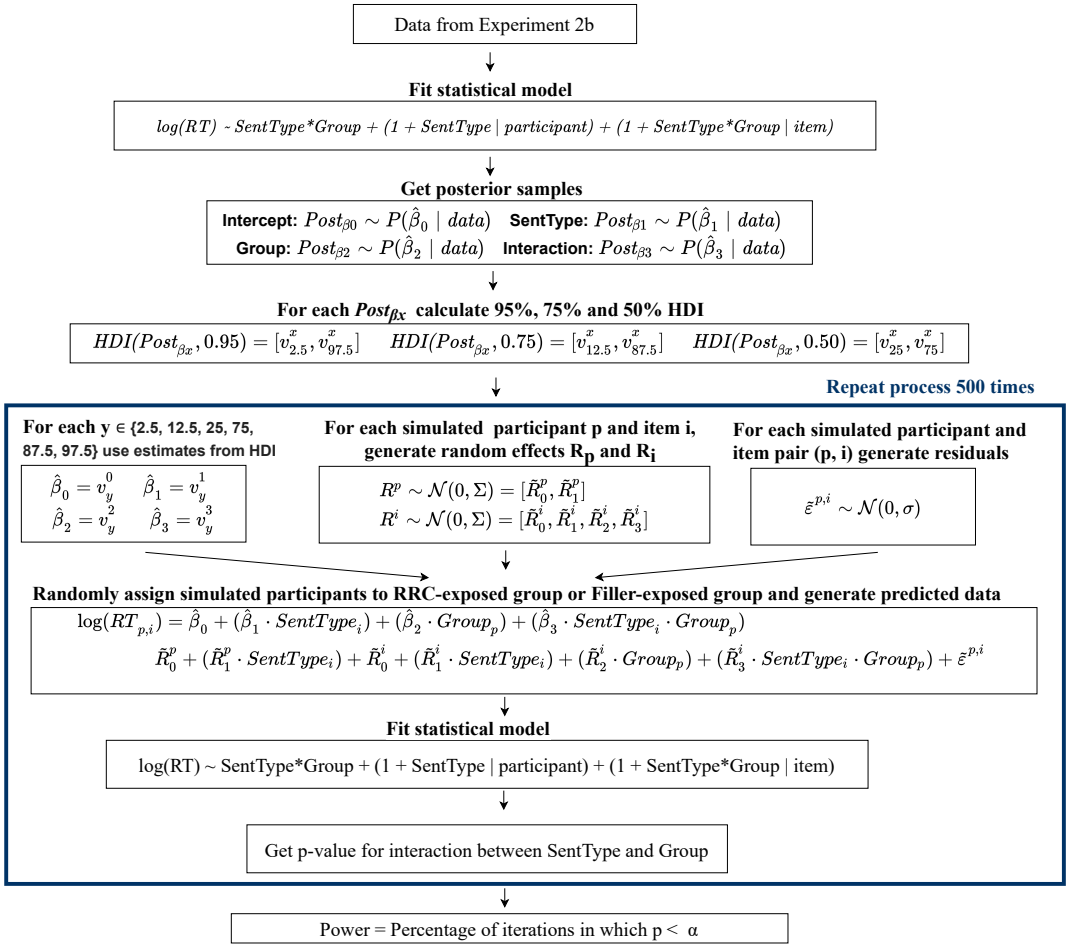


Figure 3.7: A schematic of how we calculated the power to detect a significant difference in the garden path effect between the RRC-exposed group and the Filler-exposed group for future experiments with the same design. We use the LMER notation in R for the statistical models.

the maximum likelihood estimates of all the parameters. In this analysis, by contrast, we fit a Bayesian version of the linear mixed-effects model and calculated the power based not only on the posterior mean estimates of all parameters, but also several other values of the parameters that have a range of posterior probabilities given the results of Experiment 2b. Second, in Experiment 2a we collected data from only the Filler-exposed group, we used Ω—the hypothesized ratio between the garden path effects shown by the two groups—to generate predictions for the RRC-exposed group. This hypothesized ratio was not required in the present simulations, since Experiment 2b included empirical data collected from the RRC-exposed group.

We simulated participants and items using the random effects estimated from the model fit to the results of Experiment 2b. This simulation process was identical to the prior power analysis. Then, for any given set of values of the fixed effects—the intercept (β_0), the main effect of sentence type (β_1), the main effect of group (β_2), and the interaction between these two predictors (β_3)—we generated 500 simulated RT datasets by combining the values of these fixed effects with samples from the random effects and residuals. Finally, we fit to each of these 500 datasets a new model similar to one we used to analyze the results of Experiment 2b, and calculated the proportion of simulated datasets in which β_3, the crucial interaction term, reached significance. We repeated this process separately for 200, 400 and 800 participants.

We calculated different sets of values for the fixed effects as follows. First, we fit a Bayesian version of the statistical model used in Experiment 2b. Then, we computed the highest density interval (HDI) for β_0, β_1, β_2 and β_3. An $x\%$ HDI specifies a range of values (a, b) such that $x\%$ of the posterior probability mass falls within this range. For example, if the 95% HDI for β_1 is $(0.001, 0.01)$, then

$P_{posterior}(0.001 < \beta_1 < 0.01) = 0.95$. We computed the 95%, 75% and 50% HDIs for each of the predictors and used the lower and upper bounds of these intervals as six sets of values of the fixed effects for the power analysis. For each of these six sets of values, we generated 500 datasets and calculated power as described in the previous paragraph. We also calculated power for the set of values with the posterior mean.

The Bayesian regression model we used for the power analysis differed in two ways from the shifted lognormal Bayesian regression model described above: first, we used the standard unshifted lognormal distribution and second, we used the default priors specified by the brms package (Bürkner et al., 2017): for the fixed effects, a uniform distribution over all real numbers; for the intercept, a Student's t distribution with mean 0, standard deviation 10, and 6 degrees of freedom; for the by-participant and by-item random slopes and intercepts, as well as the parameter for the residual standard deviation, a Student's t distribution with mean 0, standard deviation 10, and 3 degrees of freedom; and for the covariance matrices, LKJ Cholesky priors with $\eta = 1$. In light of the similarity between the results we obtained from the shifted distribution with informative priors and the current unshifted distribution with uninformative priors, we did not repeat our power analyses with the estimates from the shifted model.

3.5.3.2.1 Results Our power analyses indicated that if the true effect size of syntactic adaptation is the same as that observed in Experiment 2b (the posterior mean estimate), then future experiments with the design of Experiment 2b will require between 400 and 800 participants to detect a significant interaction at the $p < 0.05$ threshold with 80% power (see Figure 3.8a). If the true effect size is the highest value included in 95% HDI—1.7 times the observed effect size—then 400 participants might

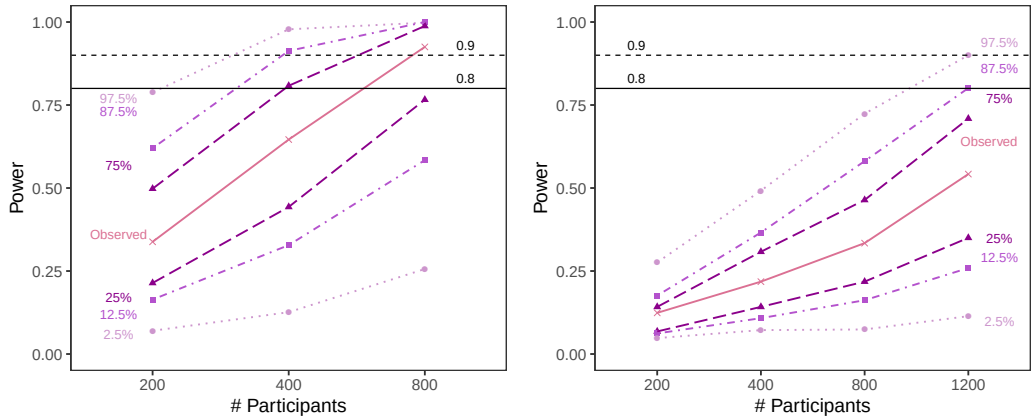


Figure 3.8: (a) Power to detect a significant interaction between group and sentence type for future studies with the same expected effect size as in Experiment 2b. (b) Power to detect a significant interaction between group and sentence type for future studies with an expected effect size of half of what was observed in Experiment 2b. Lines of the same colour and line type correspond to upper and lower bound of HDI with the same credible interval. For example, the dotted line in lightest purple reflects the upper and lower bound for the 95% HDI.

be sufficient to detect a significant interaction. On the other hand, if the true effect size is on the lower end of the 95% HDI—0.3 times the observed effect size—then even 800 participants might not be enough.[13]

3.5.3.3 How many participants would we need to detect modulations of syntactic adaptation?

The goal of Experiment 2b was to detect the *presence* of syntactic adaptation. As such, we optimized the design of that experiment to obtain the maximal possible difference in garden path effect between the two groups: in the exposure phase, Filler-exposed participants read sentences that had minimal to no structural overlap with the RRC sentences included in the test phase, whereas RRC-exposed participants were exposed to sentences that had maximal structural overlap with the test sentences.

[13]The posterior mean estimate of the interaction coefficient was 0.006 and the HDI was 0.002–0.010.

By contrast, any between-group self-paced reading experiment designed to detect *modulations* of this basic syntactic adaptation effect would likely yield smaller between-group differences than we found in Experiment 2b. Consider, for example, an experiment designed to test whether the garden path effect associated with RRCs can be diminished by repeated exposure to another type of relative clause, such as an unreduced relative clause (URC), and if so, whether the degree of adaptation differs across the two scenarios (RRC in both exposure and test, compared to URC in exposure and RRC in test). Such a hypothetical experiment would include RRC-exposed, URC-exposed and Filler-exposed groups. Any difference between RRC-exposed and URC-exposed participants is very likely be smaller than the difference between RRC-exposed and Filler-exposed groups; consequently, detecting such a modulation of syntactic adaptation would require even more participants than needed to detect its presence, as in Experiment 2b.

To estimate the power of experiments measuring such modulations of syntactic adaptation, we re-ran all the power analyses after dividing by two the upper bound and lower bound values of β_0, β_1, β_2 and β_3 described above; this expresses the assumption that modulations of the basic syntactic adaptation effect will yield smaller effect sizes than in our Experiment 2b.[14] Under these assumptions, the power analysis based on the posterior mean estimates indicated that even with 1200 participants the experiment would have only 60% power to detect a significant interaction effect at the $p < 0.05$ threshold (see Figure 3.8b). In the best case scenario, where the modulation effect size is based on the largest possible effect size contained in the 95% HDI from Experiment 2b, we would have 72% power to detect an interaction at the $p < 0.05$

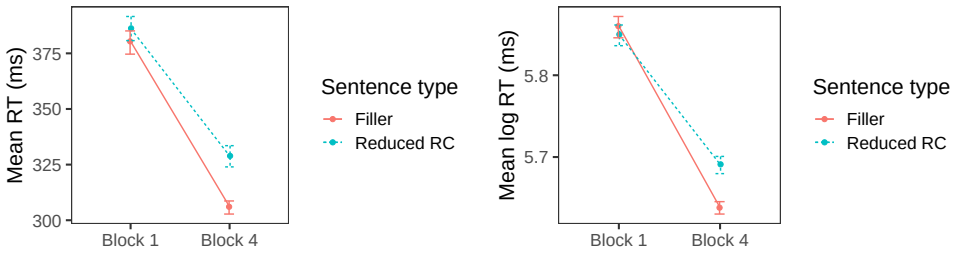
[14] Since we sampled the random effects from the original multivariate normal distributions, dividing the beta coefficients of the lower and upper bounds does not result in a decrease in the uncertainty of our estimates.

threshold with 800 participants, and 90% power with 1200 participants. In the worst case scenario, when the effect size is based on the smallest possible effect size within the same 95% HDI, we would have 7% power to detect a significant interaction with 800 participants and 11% power with 1200 participants. In other words, experiments designed to detect modulations of the syntactic adaptation effect using a between-group design could be underpowered even with as many 1200 participants.

3.5.3.4 Comparing the magnitude of task adaptation and syntactic adaptation

The reduction in the size of garden path effect is caused by task adaptation alone in the Filler-exposed group, and by both task adaptation and syntactic adaptation in the RRC-exposed group. As such, the difference in garden path effect between the two groups can be interpreted as an estimate of the effect of syntactic adaptation over and above task adaptation. In Experiment 2b, the garden path effect was 14.07 ms for the Filler-exposed group and 5.67 ms for the RRC-exposed group, as calculated from the mixed effect model estimates. This suggests that syntactic adaptation resulted in 8.4 ms decrease in the garden path effect over and above task adaptation.

This estimate has a critical limitation: it compares across two sets of participants that differ in their average reading times (see discussion of main effect of group above). To obtain an estimate of the relative magnitude of syntactic and task adaptation within participants, we focused on the RRC-exposed group, and compared the change in RTs over time between RRC sentences and filler sentences: The decrease in RTs for filler sentences is caused by task adaptation, whereas the decrease in RTs for RRC sentences is caused by a combination of task and syntactic adaptation. Therefore, if we assume that the effects of syntactic adaptation and task adaptation are additive, then we can calculate the within-participant magnitude of syntactic adaptation by



(a) **(b)**

Figure 3.9: RTs (panel a) and log RTs (panel b) averaged across sentence positions 8–10 for the RRC-exposed group in Block 1 and Block 4 for filler sentences and RRC sentences matched for RTs in Block 1. The mean RTs for all of the items in Block 1 were not greater or less than the mean RTs for all filler sentences across participants in both groups by more than 30 ms. Error bars reflect bootstrapped 95% confidence intervals.

subtracting the decrease in RTs observed in RRC sentences from that observed in filler sentences.

This within-participant comparison is again complicated by a main effect, this time the main effect of condition: because filler sentences were on average read more rapidly than RRC sentences, and because task adaptation is start-point dependent, we could not directly compare the rate of task adaptation for the RRC and filler sentences. To mitigate this, we created a subset of RRC and filler sentences that were roughly matched in difficulty: we only included a sentence if its mean RT, when averaged across all participants who read the sentence as one of the first 20 sentences, was in the range defined by the mean RT for all filler sentences in the first block $\pm$30 ms (350.9–410.9 ms). We focused on the words in positions 8–10 of both filler and RRC sentences; in RRC sentences, these are the words that make up the disambiguating region. We then averaged the RTs on these words across all the items in the subset and across all participants in the RRC-exposed group, separately when the items occurred early in the experiment (first 20 sentences) and when they occurred later in

the experiment (last 40 sentences).

If the effects of syntactic adaptation and task adaptation are additive, such that syntactic adaptation results in a decrease in RTs over and above task adaptation, then we would expect a *greater* reduction in RTs for RRC sentences than for filler sentences. Contrary to this prediction, we found that RTs decreased *less* for RRC sentences (57 ms) than for filler sentences RTs (74 ms; see Figure 3.9a). We repeated this analysis with log transformed RTs and observed qualitatively similar results (see Figure 3.9b). These surprising results suggest that on both the raw and logarithmic scale, the rate of task adaptation is lower for syntactically complex sentences than syntactically easier sentences, even when the RTs for the complex and simple sentences are matched. This poses a problem for the simplistic notion of task adaptation that we (and others) have adopted, which assumes that the effects of task adaptation and syntactic adaptation are additive and independent of each other.

3.6 General Discussion

The garden path effect observed in temporarily ambiguous sentences that are disambiguated in favor of a low-probability parse decreases over the course of a reading experiment (Fine et al., 2013; Fine and Jaeger, 2016). This finding has been interpreted as evidence that participants update their syntactic expectations to match the statistics of the environment (*syntactic adaptation*). But syntactic adaptation is not the only possible explanation for this finding: a decrease over time in the garden path effect can also be driven by the hypothesis we termed "start-point dependent task adaptation", according to which task adaptation—the decrease in RTs due to increased familiarity with the task—is greater for sentences that are read slowly when

encountered early in the experiment (“difficult sentences”) than for sentences that are initially read more rapidly (”easy sentences”). Such start-point dependent task adaptation would result in a decrease over time in the difference in reading times between easier unambiguous sentences and difficult ambiguous sentences—in other words, the garden path effect. The goal of this paper was to investigate whether syntactic adaptation results in a decrease in garden path effect over and above the decrease caused by any such start-point dependent task-adaptation.

In Experiment 1, we replicated the results of one of the experiments from (Fine and Jaeger, 2016) that have been taken as evidence for syntactic adaptation: as in their experiment, both overall reading times and the garden path effect decreased over the course of Experiment 1. We also found evidence for start-point dependent task-adaptation, suggesting that the observed decrease in garden path effect could, in theory, be entirely driven by a greater rate of task adaptation for ambiguous sentences with reduced RCs (RRC sentences) than unambiguous ones with unreduced RCs (URC sentences).

The main experiment of the paper was Experiment 2b, whose goal was to detect syntactic adaptation over and above task adaptation. This experiment compared the garden path effects in two groups of participants: one exposed to filler sentences only (Filler-exposed group), and the other exposed to both filler and RRC sentences (RRC-exposed group). Following the exposure phase, both groups read RRC and URC sentences. In the Filler-exposed group, only task adaptation was possible, whereas in the RRC-exposed group both task and syntactic adaptation were possible.

Before running Experiment 2b, we ran a preliminary experiment, Experiment 2a, in which we collected data from Filler-exposed participants only, and used it to estimate

the number of participants to run in Experiment 2b. We estimated that the number of participants required to reliably detect a significant difference in garden path effect between the two groups can be as high as 800. Consequently, in Experiment 2b, we collected data from 828 participants, 642 of whom were included in the analyses.

Experiment 2b showed that after the exposure phase, the garden path effect for the RRC-exposed group was diminished compared to that of the Filler-exposed group. Since both groups were exposed to the same number of sentences during the exposure phase, the difference in garden path effect between the groups cannot be completely explained by task adaptation, and has to be driven by the difference in the types of sentences that the participants were exposed to (i.e. RRC sentences vs. filler sentences). As such, these results support the hypothesis that syntactic adaptation causes a decrease in the garden path effect over and above the decrease caused by task-adaptation.

We next conducted a Bayesian analysis to estimate the range of effect sizes that are plausible given our data, and used those to estimate the power required to detect an effect in future studies with the same experimental design as Experiment 2b. This power analysis indicated that if the true effect size is equal to the effect observed in our experiment, then future experiments would require between 400 and 800 participants to have 80% power to detect the difference in garden path effect between groups. If the true effect size is smaller than that observed in our experiment, but still within the 95% credible interval given our results, then future experiments with the same design are likely to be underpowered with even 800 participants. Finally, we estimated the power to detect an effect in future between-group studies with similar experimental setup as Experiment 2b aimed at investigating how syntactic adaptation interacts with other

factors. Under the assumption that such subtler effects result in an effect size half as large as in Experiment 2b, we found that these experiments could be underpowered even with as many as 1200 participants.

3.6.1 Why are so many participants required to reliably detect effects of syntactic adaptation in self-paced reading?

We discuss two possible answers to this question: first, that a decrease in garden path effect in a self-paced reading experiment is not an ideal dependent measure if the goal is to detect syntactic adaptation; and second, that syntactic adaptation results in very small and hard-to-measure changes to readers' expectations, more generally.

3.6.1.1 Explanation 1: Decrease in garden path effect in self-paced reading is a dependent measure that is ill-suited for studying syntactic adaptation.

It is possible that syntactic adaptation can, in principle, be reliably detected with fewer participants in a between-group design than our power analysis suggests, but that self-paced reading is not an ideal paradigm to do so. As discussed earlier, task adaptation in this paradigm is start-point dependent; this leads to a compression over time of the difference in RTs between "easy" and "difficult" sentences, independently of any syntactic properties of those sentences. This compression causes a reduction in garden path effect. The high rates of task adaptation in self-paced reading therefore lead to smaller garden-path effects overall in the later parts of the experiment. This in turn results in a smaller absolute between-group differences in garden path effect. Since smaller effect sizes are often accompanied by lower power, more participants are likely to be required to detect effects of syntactic adaptation.[15]

[15] In principle, it is possible for power to stay the same as the effect size decreases, if the variability in the data also decreases along with the effect size. To test this, we refit the statistical model from

This explanation points to two alternative methods of measuring syntactic adaptation that might result in larger effects: first, using a dependent measure that is not confounded with task adaptation; second, using a paradigm where task adaptation is not start-point dependent. It is unclear whether the latter method is currently feasible, since we are unaware of paradigms where task adaptation has been demonstrated to be start-point independent. However, a reviewer pointed out that there is indeed a dependent measure of syntactic adaptation that is not confounded with task adaptation — an *increase* in the garden path effect for sentences disambiguated in favor of the main verb reading, as in (6):

(6) The evil genie served the golden figs before going into a trance.

Since task adaptation results in a *decrease* in garden path effect, it would be possible to circumvent the loss in power due to task-adaptation even in self-paced reading studies, if we used the increase in garden path effect as a dependent measure. A potential concern with using the *increase* in garden path effect as a dependent measure is that, under the expectation adaptation account, after n observations, there is a greater change in surprisal for unexpected structures (reduced RC reading) than for sentences with expected structure (MV reading) (Jaeger, Bushong, and Burchill, 2019). Therefore, detecting an increase in the garden path effect for sentences with a MV reading can be much more challenging than detecting a decrease in the garden path effect for sentences with reduced RC reading. Further simulations and experiments are required to investigate whether the advantage of using the increase in garden path

Analysis 1.1 separately on the first two and the last two blocks of Experiment 1. If the variability in the data decreased along with the effect size, we would expect both the estimate of garden path effect and the standard error in the last two blocks to be lower than in the first two. In contrast to this prediction, we found that while the estimate of garden path effect decreased (from 0.044 in the first block to 0.007 in the last block), the standard error of the estimates remained the same (0.007 in both blocks).

effects as a dependent measure (it is not correlated with task-adaptation) outweighs the disadvantage (it is predicted to have a smaller effect size).

3.6.1.2 Explanation 2: Syntactic adaptation results in extremely small changes to our expectations.

An alternative explanation, which is also consistent with our results, is that exposure to sentences with unexpected structures in the context of an experiment results in extremely small changes to our expectations. If that is the case, syntactic adaptation may be difficult to observe irrespective of the paradigm or dependent measure we use. If the true effect size of syntactic adaptation is indeed very small, then this raises a broader question: what constitutes a psychologically meaningful effect size? The answer to this question can vary depending on the goals of the research program. If the goal is to apply the findings from the syntactic adaptation literature in a practical context (e.g., in education), then extremely small effect sizes might not be meaningful. On the other hand, if the goal is to build a theory on the basis of syntactic adaptation, then extremely small effect sizes might be meaningful, but not practical to study. Finally, if the goal is to only use syntactic adaptation to verify one of the predictions of a larger theoretical framework, then extremely small effect sizes can be both meaningful and practical.

3.6.2 What properties of RRC sentences are participants adapting to?

Experiment 2b indicated that participants in the RRC-exposed group adapted to some property of the RRC sentences they were exposed to, but did not isolate the property (or properties) of the RRC sentences to which participants were adapting. Following

previous papers on syntactic adaptation, we assumed that participants updated their expectations about an abstract grammar rule such as "the subject of the sentence is modified by a reduced relative clause". However, it is also possible that participants were adapting to an accidental property of RRC sentences included in the experiment, such as the fact that the seventh word of the sentence was always a verb; or that they were adapting their parsing strategies to the large number of temporarily ambiguous sentences included in the experiment, for example by maintaining a larger number of potential parses for each sentence (Jurafsky, 1996).

In future work, these possibilities can be distinguished by measuring the magnitude of syntactic adaptation for sentences with varying properties. For example, if syntactic adaptation is weaker when the verbs in the exposure sentence occur in varying positions than when they occur in the same position, we can conclude that participants were adapting to the position of the verb in the sentence. Similarly, if syntactic adaptation is stronger when the exposure phase contains other types of garden path sentences (e.g., *When Anna bathed the baby spit up*) than when it contains filler sentences only, we can conclude that participants were adapting to the prevalence of temporarily ambiguous sentences in the experiment. As discussed earlier, the power of such experiments, which are designed to measure modulations of the syntactic adaptation effect, is likely to be relatively low in self-paced reading studies with designs similar to Experiment 2b.

3.7 Conclusion

This study provided evidence for rapid syntactic adaptation in self-paced reading studies using a between-group experimental setup. At the same time, hundreds

of participants were required to detect a syntactic adaptation over and above the substantially stronger effect of adaptation to the self-paced reading task. Power analyses indicated that experiments with a similar between-group design whose goal is to study factors that modulate this effect, such as the particular syntactic properties that participants are able to adapt to, will likely require even more participants. We conclude that theoretical questions about syntactic adaptation are likely to be more fruitfully addressed using experimental paradigms that are not confounded with task adaptation, or paradigms in which task adaptation is not start-point dependent (if such paradigms exist).

3.8 Acknowledgments

We would like to thank Nicholas Douglass and Brian Leonard for assistance with stimulus creation and data collection respectively. We would also like to thank Duane Watson and the two anonymous reviewers for their extensive constructive feedback which helped us significantly improve the paper. Finally, we would like to thank Brian Dillon, the members of the Computation and Psycholinguistics lab at Johns Hopkins University, and audiences at the 2019 and 2020 CUNY Conferences on Human Sentence Processing for their helpful discussion of earlier versions of this work.

References

Mitchell, Don C., Fernando Cuetos, Martin M. B. Corley, and Marc Brysbaert (1995). "Exposure-Based Models of Human Parsing: Evidence for the Use of Coarse-Grained (Nonlexical) Statistical Records". In: *Journal of Psycholinguistic Research* 24.6, pp. 469–488.

Romberg, Alexa R. and Jenny R. Saffran (2010). "Statistical learning and language acquisition". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.6, pp. 906–914.

MacDonald, Maryellen C., Neal J. Pearlmutter, and Mark S. Seidenberg (1994). "The lexical nature of syntactic ambiguity resolution." In: *Psychological Review* 101.4, pp. 676–703. URL: http://dx.doi.org/10.1037/0033-295X.101.4.676.

Trueswell, John C. (1996). "The Role of Lexical Frequency in Syntactic Ambiguity Resolution". In: *Journal of Memory and Language* 35.4, pp. 566–585. URL: https://doi.org/10.1006/jmla.1996.0030.

Anderson, John R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Fine, Alex B., T. Florian Jaeger, Thomas A. Farmer, and Ting Qian (2013). "Rapid Expectation Adaptation during Syntactic Comprehension". In: *PLoS One* 8.10, e77661. URL: https://doi.org/10.1371/journal.pone.0077661.

Wells, Justine B., Morten H. Christiansen, David S. Race, Daniel J. Acheson, and Maryellen C. MacDonald (2009). "Experience and sentence processing: Statistical learning and relative clause comprehension". In: *Cognitive Psychology* 58, pp. 250–271.

Liversedge, Simon P, Kevin B Paterson, and Emma L Clayes (2002). "The influence of only on syntactic processing of "long" relative clause sentences". In: *The Quarterly Journal of Experimental Psychology Section A* 55.1, pp. 225–240.

Clifton Jr, Charles, Matthew J Traxler, Mohamed Taha Mohamed, Rihana S Williams, Robin K Morris, and Keith Rayner (2003). "The use of thematic role information

in parsing: Syntactic processing autonomy revisited". In: *Journal of Memory and Language* 49.3, pp. 317–334.

Kemper, Susan, Angela Crow, and Karen Kemtes (2004). "Eye-fixation patterns of high-and low-span young and older adults: down the garden path and back again." In: *Psychology and Aging* 19.1, p. 157.

Hale, John (2001). "A Probabilistic Earley Parser As a Psycholinguistic Model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Pittsburgh, Pennsylvania: Association for Computational Linguistics, pp. 1–8. DOI: 10.3115/1073336.1073357. URL: https://doi.org/10.3115/1073336.1073357.

Ehrlich, S.F. and K. Rayner (1981). "Contextual effects on word perception and eye movements during reading". In: *Journal of Verbal Learning and Verbal Behavior* 20.6, pp. 641–655.

Smith, Nathaniel J and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128.3, pp. 302–319.

Fine, Alex B. and T. Florian Jaeger (2016). "The role of verb repetition in cumulative structural priming in comprehension". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42.9, pp. 1362–1376. URL: http://dx.doi.org/10.1037/xlm0000236.

Stack, Caoimhe M. Harrington, Ariel N. James, and Duane G. Watson (2018). "A failure to replicate rapid syntactic adaptation in comprehension". In: *Memory and Cognition* 46.6. DOI: 10.3758/s13421-018-0808-6.

Jaeger, TF, WR Bushong, and Z Burchill (2019). *Strong evidence for expectation adaptation during language understanding, not a replication failure. A reply to Harrington Stack, James, and Watson (2018)*.

Drummond, Alex (2016). *Ibex Farm*. https://github.com/addrummond/ibex.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

York, Richard (2012). "Residualization is not the answer: Rethinking how to address multicollinearity". In: *Social Science Research* 41.6, pp. 1379–1386.

Heathcote, Andrew, Scott Brown, and D.J.K Mewhort (2000). "The power law repealed: The case for an exponential law of practice". In: *Psychonomic Bulletin & Review* 7.2, pp. 185–207.

Malone, Avery and Gail Mauner (2018). "What Do Readers Adapt to in Syntactic Adaptation?" In: *Poster session presented at the 31st Annunal CUNY Sentence Processing Conference*. Davis, CA, USA.

Roland, Douglas and Daniel Jurafsky (2002). "Verb sense and verb subcategorization probabilities". In: *The lexical basis of sentence processing: Formal, computational, and experimental issues*, pp. 325–346.

Fine, Alex B, Ting Qian, T Florian Jaeger, and Robert A Jacobs (2010). "Is there syntactic adaptation in language comprehension?" In: *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, pp. 18–26.

Leonard, Brian (2019). *mTurk-Microbatcher*. https://github.com/jhupsycholing/mturk-microbatcher.

Rouder, Jeffrey N (2005). "Are unshifted distributional models appropriate for response time?" In: *Psychometrika* 70.2, p. 377.

Schad, Daniel J, Michael Betancourt, and Shravan Vasishth (2019). "Toward a principled Bayesian workflow in cognitive science". In: *arXiv preprint arXiv:1904.12765*.

Bürkner, Paul-Christian et al. (2017). "brms: An R package for Bayesian multilevel models using Stan". In: *Journal of Statistical Software* 80.1, pp. 1–28.

Jurafsky, D. (1996). "A probabilistic model of lexical and syntactic access and disambiguation". In: *Cognitive Science* 20.2, pp. 137–194.

Chapter 4

What is the system of rules that governs the incremental structures that neural networks build?

This chapter was previously published as:

Prasad, G., van Schijndel, M & Linzen, T. (2019). Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

4.1 Introduction

Neural networks trained on text alone, without explicit syntactic supervision, have been surprisingly successful in tasks that require sensitivity to sentence structure. The difficulty of interpreting the learned neural representations that underlie this success has motivated a range of analysis techniques, including diagnostic classifiers Giulianelli et al. (2018), Conneau et al. (2018), and Shi, Padhi, and Knight (2016), visualization of individual neuron activations Kádár, Chrupała, and Alishahi (2017)

and Qian, Qiu, and Huang (2016), ablation of individual neurons or sets of neurons Lakretz et al. (2019) and behavioral tests of generalization to infrequent or held out syntactic structures Linzen, Dupoux, and Goldberg (2016), Weber, Shekhar, and Balasubramanian (2018), and McCoy, Frank, and Linzen (2018); for reviews, see Belinkov and Glass (2019) and Alishahi, Chrupała, and Linzen (2019).

This paper expands the toolkit of neural network analysis techniques by drawing on the **syntactic priming** paradigm, a central tool in psycholinguistics for analyzing human syntactic representations Bock (1986). This paradigm is based on the empirical finding that people tend to reuse syntactic structures that they have recently produced or encountered. For example, English provides two roughly equivalent ways to express a transfer event:

(1) a. The boy threw the ball to the dog.

b. The boy threw the dog the ball.

When readers encounter one of these variants in the text more frequently than the other, they expect that future transfer events will more likely be expressed using the frequent construction than the infrequent one. For example, after reading sentences like (1a) (the **prime**), readers expect sentences like (2a), which shares syntactic structure with the prime, to occur with a greater likelihood than the alternative variant like (2b) which does not Wells et al. (2009).[1]

(2) a. The lawyer sent the letter to the client.

b. The lawyer sent the client the letter.

[1]Wells et al. (2009) measured priming effects for relative clauses, not dative constructions. For work on priming in production with dative constructions, see Kaschak, Kutta, and Jones (2011).

We use the priming paradigm to analyze neural network language models (LMs), systems that define a probability distribution over the n^{th} word of a sentence given its first $n-1$ words. Building on paradigms that determine whether the LM's expectations are consistent with the syntactic structure of the sentence Linzen, Dupoux, and Goldberg (2016), we measure the extent to which a LM's expectation for a specific syntactic structure is affected by recent experience with related structures. We prime a fully trained model with a structure by adapting it to a small number of sentences containing that structure (Schijndel and Linzen, 2018). We then measure the change in surprisal (negative log probability) after adaptation when the LM is tested either on sentences with the same structure or sentences with different but related structures. The degree to which one structure primes another provides a graded similarity metric between the model's representations of those structures (cf. Branigan and Pickering (2017)), which allows us to investigate how the representations of sentences with these structures are organized.

As a case study, we applied this technique to investigate how recurrent neural network (RNN) LMs represent sentences with relative clauses (RCs). We found that the representations of these sentences are organized in a linguistically interpretable manner: sentences with a particular type of RC were most similar to other sentences with the same type of RC in the LMs' representation space. Furthermore, sentences with different types of RCs were more similar to each other than sentences without RCs. We demonstrate that the similarity between sentences was not driven merely by specific words that appeared in the sentence, suggesting that the LMs tracked abstract properties of the sentence. This ability to track abstract properties decreased as the training corpus size increased. Finally, we tested the hypothesis that LMs' accuracy

Abstract structure	Example
Unreduced Object RC	The conspiracy that the employee welcomed divided the beautiful country.
Reduced Object RC	The conspiracy the employee welcomed divided the beautiful country.
Unreduced Passive RC	The conspiracy that was welcomed by the employee divided the beautiful country.
Reduced Passive RC	The conspiracy welcomed by the employee divided the beautiful country.
Active Subject RC	The employee that welcomed the conspiracy quickly searched the buildings.
PS/ORC-matched Coordination	The conspiracy welcomed the employee and divided the beautiful country.
ASRC-matched Coordination	The employee welcomed the conspiracy and quickly searched the buildings.

Table 4.1: Examples of sentences generated using templates containing the seven abstract structures we analyzed (optional elements, which only occur in a subset of the examples, are indicated in grey).

on agreement prediction (Marvin and Linzen, 2018) would increase with the LMs' ability to track more abstract properties of the sentence, but did not find evidence for this hypothesis.

4.2 Background

4.2.1 Syntactic predictions in neural LMs

We build on paradigms that use LM probability estimates for words in a given context as a measure of the model's sensitivity to the syntactic structure of the sentence Linzen, Dupoux, and Goldberg (2016), Gulordava et al. (2018), and Marvin and Linzen (2018). If a language model assigns a higher probability to a verb form that agrees in number with the subject (*the boy... writes*) than a verb form that does not (*the boy... write*), we can infer that the model encodes information about the agreement features of nouns and verbs (that is, the difference between singular and plural) and has correctly identified the subject that corresponds to this verb. This reasoning has been extended beyond subject-verb agreement to study whether the predictions of neural LMs are sensitive to a range of other syntactic dependencies, including negative polarity items Jumelet and Hupkes (2018), filler-gap dependencies Wilcox et al. (2018) and reflexive

pronoun binding Futrell et al. (2019).

4.2.2 Syntactic priming in humans

Syntactic priming has been used to study whether the representations of two sentences have shared structure. For example, (1a) (repeated below as (3)) shares the structure VP $\rightarrow$ V NP PP with (4a) but not (4b).

(3) The boy threw the ball to the dog.

(4) a. The renowned chef made some wonderful pasta for the guest.

b. The renowned chef made the guest some wonderful pasta.

If (3) primes (4a) more than it primes (4b), we can infer that the representations of (3) are more similar to that of (4a) than to that of (4b). Since (4b) and (4a) differ only in their structure, this difference in similarity must be driven by structural information in the representations of the sentences (for reviews, see Mahowald et al. (2016) and Tooley and Traxler (2010)).

Although priming studies have traditionally measured the priming effect on the sentence immediately following the prime, more recent studies have demonstrated that the effects of syntactic priming can be cumulative and long-lasting: sentences with a shared structure S_X become progressively easier to process when preceded by n sentences with the same structure S_X than when preceded by n sentences with a different structure S_Y Kaschak, Kutta, and Jones (2011) and Wells et al. (2009).[2] In conjunction with the finding that words that are consistent with a probable syntactic parse are easier to process than words consistent with less probable parses Hale (2001)

[2]In studies looking at non-cumulative priming, $n = 1$.

and Levy (2008), the increased ease of processing in cumulative priming studies can be interpreted as evidence that, with increased exposure to a structure, participants begin to expect that structure with a greater probability Chang, Dell, and Bock (2006).

Cumulative priming allows us to study how sentences are related to each other in the human (or LM) representation space in the same way that non-cumulative priming does: when participants (or LMs) are exposed to sentences with structure S_X, if there is a greater decrease in surprisal when they are tested on other sentences with S_X than when they are tested on other sentences with S_Y, we can infer that the representations of sentences with S_X are more similar to each other than to the representations of sentences with S_Y.

4.2.3 LM adaptation as cumulative priming

Schijndel and Linzen (2018) modeled cumulative priming in recurrent neural networks (RNNs) by adapting fully trained RNN LMs to new stimuli — i.e. taking a fully trained RNN LM and continuing to train it on a small set of sentences (cf.,Grave, Joulin, and Usunier (2017), Krause et al. (2017), and Chowdhury and Zamparelli (2019)). They demonstrated that when an RNN LM was adapted to a small number of sentences with a shared syntactic structure, the surprisal for novel sentences with that structure decreased, enabling them to infer that the LM's representations of sentences contained information about that structure.

4.3 Similarity between syntactic structures in RNN LM representational space

Following the assumptions in Section 4.2.2, we define a similarity metric between two structures S_X and S_Y in an LM's representation space by adapting the LM to sentences with S_X and measuring the change in surprisal for sentences with S_Y — i.e. measuring to what extent sentences with S_X prime sentences with S_Y. We use the notation $\mathbb{A}(Y \mid X)$ to refer to this change in surprisal[3], where X and Y are non-lexically-overlapping sets of sentences whose members share the structures S_X and S_Y respectively. If we assume that S_X and S_Y are similar to each other in the LM's representation space, then $\mathbb{A}(Y \mid X) > 0$ — i.e., encountering sentences with S_X causes the LM to assign a higher probability to sentences with S_Y. On the other hand, if we assume that S_X and S_Y are unrelated to each other, then $\mathbb{A}(Y \mid X) = 0$ — i.e., encountering sentences with S_X does not cause the LM to change its probability for sentences with S_Y.

4.4 Experimental setup

4.4.1 Syntactic structures

We analyzed five types of RCs. In an active subject RC, the gap is in the subject position of the embedded clause:[4]

(5) My cousin that _ liked the book ...

[3] $\mathbb{A}$ is shorthand for adaptation.

[4] We illustrate the location of the gap with underscores here, but the underscores were not included in the LM's input.

In a passive subject RC (*passive RCs*), the gap is in the subject position of the embedded clause, and the embedded verb is passive. In English, passive RCs can be unreduced (6a) or reduced (6b):

(6) a. The book that _ was liked by my cousin ...

b. The book _ liked by my cousin ...

In an object RC the gap is in the object position of the embedded clause. In English, object RCs can be unreduced (7a) or reduced (7b):

(7) a. The book that my cousin liked _ ...

b. The book my cousin liked _ ...

Finally, we also included two additional conditions with verb coordination: one with nearly identical word order and lexical content as active subject RCs ((8); ASRC-matched Coordination), and another with nearly identical word order and lexical content as passive RCs and object RCs ((9); PS/ORC-matched Coordination).[5]

(8) My cousin liked the book and ...

(9) The book liked my cousin and ...

These conditions enable us to measure whether sentences with different types of RCs are more similar to each other in an LM's representation space than they are to lexically matched sentences without RCs.

[5]In order to maintain the same word order as in object and passive RCs, the subject of the coordinated verb phrases is an NP that tends to fill the object position in other sentences (e.g, "the equation"). Therefore, many of the sentences in this condition are implausible (e.g., "The equation reviewed the physicists and challenged the method.")

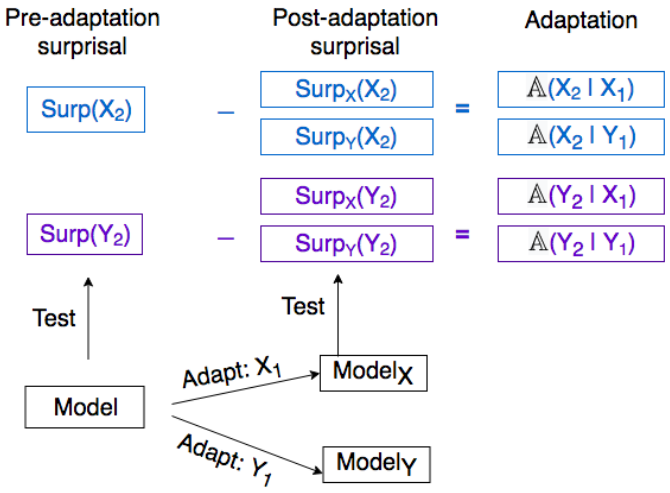


Figure 4.1: A schematic for calculating the similarity between two structures S_X and S_Y in an LM's representation space. X_1, X_2 and Y_1, Y_2 are non-lexically-overlapping sets of sentences with S_X and S_Y respectively. $Model_X$ and $Model_Y$ refer to versions of a fully trained model that have been adapted to either X_1 or Y_1 respectively. $Surp_X()$ and $Surp_Y()$ are functions that return the surprisal of sentences for $Model_X$ and $Model_Y$.

4.4.2 Adaptation and test sets

We generated sentences from seven templates, one for each of the syntactic structures of interest. The slots were filled with 223 verbs, 164 nouns, 24 adverbs and 78 adjectives such that the semantic plausibility of the combination of nouns, verbs, adverbs and adjectives was ensured. The seven variants of every sentence had nearly identical lexical items (see Table 4.1).[6] We used these templates to generate five experimental lists — each list comprised of a pair of adaptation and test sets with minimal lexical overlap between them (only function words and some modifiers were shared). Each adaptation set contained 20 sentences and each test set contained 50.

[6]Since the main verb of the sentence was constrained to be semantically plausible with the subject of the sentence, it often varied between active subject RC and ASRC-matched coordination on the one had and all other conditions on the other.

In order to infer that any decrease in surprisal is caused by adaptation to an abstract syntactic structure, we need to ensure that the models are not adapting to properties of the sentence that are unrelated to the abstract structure of interest. Consider a LM adapted to (10) and tested on (11):

(10) The conspiracy that the employee welcomed divided the country.

(11) The proposal that the receptionist managed shocked the CEO.

When the LM is adapted to sentences such as (10), it could adjust its expectations about several properties of the sentence, some more linguistically interesting than others. For instance, it could learn that there are three determiners in the sentence, that the third word of the sentence is *that*, that sentences have nine words, that every verb is preceded by a noun, and so on and so forth. If there is a decrease in surprisal when a model is adapted to (10) and tested on (11), it is unclear if this is because the model learned to expect object relative clauses or if it learned to expect any of the other mentioned properties.

To minimize the likelihood that the adaptation effects are driven by irrelevant properties of the sentence, we introduced several sources of variability to our templates: nouns could either be singular or plural, noun phrases could be optionally modified by an adjective, adjectives were optionally modified with an intensifier and verb phrases were optionally modified with adverbs which could occur either pre-verbally or post-verbally (details in the Supplementary Materials).[7]

[7]The templates and code for all the analyses along with the data can be found on GitHub: https://github.com/grushaprasad/RNN-Priming

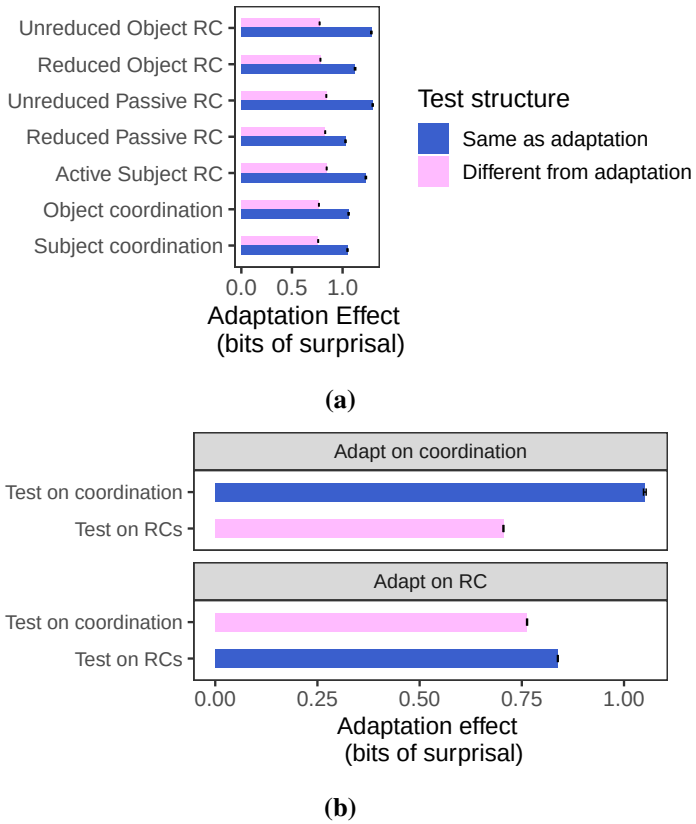


Figure 4.2: The adaptation effect averaged across all 75 models when (a) they were adapted to each of the structures and tested on either the same structure (blue, bottom) or different structure (pink, top) and (b) they were adapted to RCs and tested on non-RCs or vice versa (pink bars); or when they were adapted to RCs or non-RCs and tested on other RCs or and non-RCs respectively (blue bars). Greater values indicate more similarity between adaptation and test structures. Error bars reflect 95% CIs.

4.4.3 Models

We used 75 of the LSTM language models trained by Schijndel, Mueller, and Linzen (2019); these LMs varied in the number of hidden units per layer (100, 200, 400, 800, 1600) and the number of tokens they were trained on (2 million, 10 million or 20 million). For each training corpus size, Schijndel and Linzen trained models on five disjoint subsets of the WikiText-103 corpus, to ensure that the results generalized across different training sets.

4.4.4 Calculating the adaptation effect (AE)

For every structure, we computed the similarity between that structure and every other structure (including itself) as described in Section 4.3. This process is schematized in Figure 4.1. The surprisal values were averaged across the entire sentence.[8]

We found that $\mathbb{A}(B \mid A)$ was proportional to the surprisal of B prior to adaptation (see Supplementary Materials). As a consequence, for three structures X, Y and Z, $\mathbb{A}(Y \mid X)$ could be greater than $\mathbb{A}(Z \mid X)$ merely because Y was a more surprising structure to begin with than Z. In order to remove this confound, we first fit a linear regression model predicting $\mathbb{A}(Y \mid X)$ from the surprisal of Y prior to adaptation ($Surp(Y)$):

$$\mathbb{A}(Y \mid X) = \beta_0 + \beta_1 Surp(Y) + \epsilon$$

We then regressed out the linear relationship between $\mathbb{A}(Y \mid X)$ and $Surp(Y)$ as follows:

[8] Unknown words were excluded from this average.

$$AE(Y \mid X) = \mathbb{A}(Y \mid X) - \beta_1 Surp(Y)$$

$$= \beta_0 + \epsilon$$

Since $Surp(Y)$ was centered around its mean, β_0 reflects the mean of $\mathbb{A}(Y \mid X)$ when $Surp(Y)$ is equal to the mean surprisal of all sentences prior to adaptation. The term ϵ reflects any variance in $\mathbb{A}(Y \mid X)$ that is not predicted by $Surp(Y)$. By summing these two terms together, $AE(Y \mid X)$ reflects the change in surprisal for Y after adapting to X that is independent of $Surp(Y)$.

4.4.5 Statistical analyses

We used linear mixed effects models Pinheiro, Bates, et al. (2000) to test for statistical significance; all of the results reported below were highly significant. Details about the statistical analyses can be found in the Supplementary Materials.

4.5 Results

4.5.1 Validating AE as a similarity metric

As discussed in Section 4.2.3, under the adaptation-as-priming paradigm, we would expect sentences that share the same specific structure to be more similar to each other than lexically matched sentences that do not share the structure.[9] In other words, if X_1 and X_2 are non-lexically-overlapping sets of sentences with shared structure S_X, and Y_2 is a set of sentences with structure S_Y, but is lexically matched with X_2, then we would expect $AE(X_2 \mid X_1) > AE(Y_2 \mid X_1)$. We found this prediction to be true for all of our seven structures (Figure 4.2a), thus validating our similarity metric.

[9] By lexically matched we mean that all content words were shared between sentences.

4.5.2 Similarity between sentences with different types of VP coordination

Our two coordination conditions were structurally identical to each other but varied in their semantic plausibility — the sentences in PS/ORC-matched coordination condition were often semantically implausible whereas sentences in ASRC-matched condition were always semantically plausible (see footnote 5). If sentences that were structurally similar were close together irrespective of semantic plausibility, then we expect sentences with coordination to be more similar to each other than lexically matched sentences with RCs. Consistent with this prediction, the adaptation effect for models adapted to one type of coordination was greater when the models were tested on sentences with the other type of coordination than when they were tested on sentences with RCs (top panel of Figure 4.2b).

4.5.3 Similarity between sentences with different types of RCs

Unlike sentences with coordination, sentences with different types of RCs differ from each other at a surface level (see Table 4.1). However, at a more abstract level they all share a common property: a gap. If the RNN LMs were keeping track of whether or not a sentence contained a gap, we would expect sentences with different types of RCs to be more similar to each other in the RNN LMs' representation space than lexically matched sentences without a gap. In other words, if RC_X and RC_Y are two different types of RCs and $Coord_Y$ is a sentence with verb coordination lexically matched with RC_Y, then we would expect $AE(RC_Y \mid RC_X) > AE(Coord_Y \mid RC_X)$.

Consistent with this prediction, the adaptation effect for models adapted to RCs was greater when they were tested on sentences with other types of RCs than when

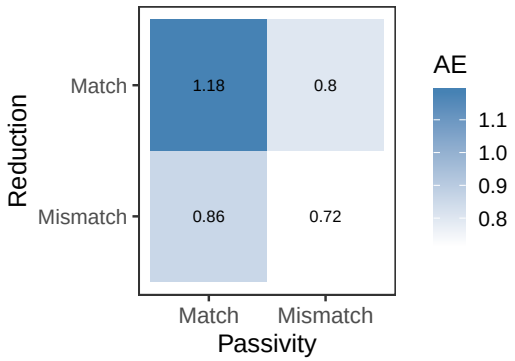


Figure 4.3: The adaptation effect when models adapted to sentences with reduced and unreduced RCs are tested on sentences that match only in reduction (top right), match only in passivity (bottom right), match in both reduction and passivity (top left) or sentences that match in neither (bottom right).

they were tested on sentences with coordination (bottom panel of Figure 4.2b). This suggests that the LMs do keep track of whether or not a sentence contains a gap, even though this property is not overtly indicated by a lexical item that is shared across all types of RCs.

4.5.4 Similarity between sentences belonging to different sub-classes of RCs

The different types of RCs we tested can be divided into sub-classes based on at least two linguistically interpretable features: reduction and passivity. Reduction distinguishes reduced passive and object RCs on the one hand from unreduced passive and object RCs on the other. Passivity distinguishes reduced and unreduced passive RCs on the one hand from reduced and unreduced object RCs on the other. The LMs could be tracking either, both or none of these features.

We probed whether the LMs track these features by comparing the similarity between sentences that share one feature but not the other, with the similarity between

sentences that share neither feature. If the adaptation effect is greater when there is a match in one feature than when there is a match in neither of the features, we can infer that the LMs track whether sentences have that feature. We found that the LMs track both of these features (Figure 4.3).

Additionally, we probed which of the features contributes more towards the similarity between sentences by comparing the similarity between sentences that match only in passivity with sentences that match only in reduction. When the adaptation and test sets matched only in passivity, the adaptation effect was slightly (but significantly) greater than when the adaptation and test sets matched only in reduction (Figure 4.3). In other words, in the LMs' representation space, (12) is more similar to (13) than it is to (14), suggesting that passivity contributes more towards the similarity between sentences than reduction.

(12) The conspiracy the employee welcomed divided the country.

(13) The conspiracy that the employee welcomed divided the country.

(14) The conspiracy welcomed by the employee divided the country.

This result is both intuitive and linguistically interpretable — the edit distance between reduced and unreduced RCs is smaller than the that between object and passive RCs; the syntax tree for (12) is also more similar to (13) than it is to (14).

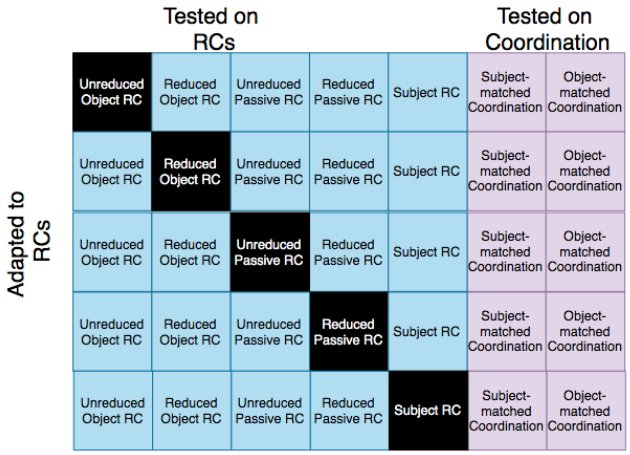


Figure 4.4: A schematic of how $\mathbb{D}(RC, \neg RC)$ is calculated. For any given row, the black square indicates the specific structure the models were adapted to, the blue squares indicate other structures that belong to the same linguistically defined class as the black square and the pink squares indicate the structures that do not belong to this linguistically defined class. In calculating the distance, we first calculated the proportion between the mean adaptation effect for the blue squares and the mean adaptation effect for pink squares for each row. We then averaged across the proportion for each row to arrive at one number.

4.5.5 What properties of sentences drive the similarity between them?

Our analyses so far have demonstrated that sentences that belong to linguistically interpretable classes (e.g., sentences that match in reduction) are more similar to each other in the LMs' representation space than they are to sentences that do not belong to those classes (e.g., sentences that do not match in reduction). However, it is unclear what properties of the sentences are driving this similarity between members of the class. For almost all of the linguistically interpretable classes we considered, all sentences belonging to a class shared at least some, if not all, function words. The only exception was the class of all RCs, where the property shared by all sentences in this class (the presence of a gap) was not overtly observable. Therefore, it is possible that the similarity between members of most of the classes we tested was being driven

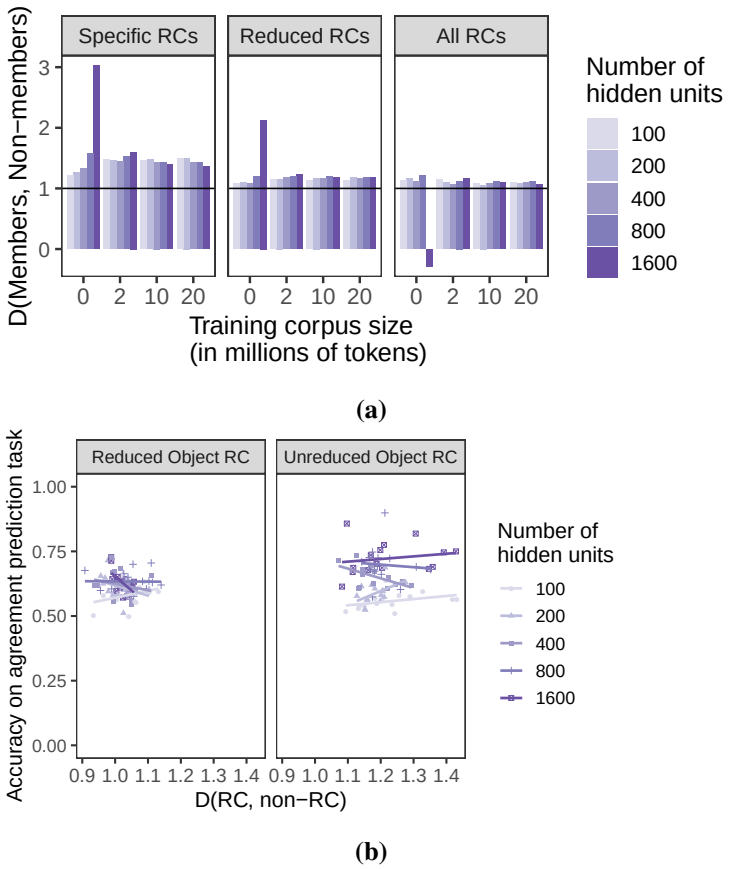


Figure 4.5: (a) Effect of hidden layer size and corpus size on the distance between sentences with specific RCs and sentences without (left), between sentences that match in reduction and sentences that do not (middle) and between sentences with RCs and sentences without (right). The solid black line indicates the point at which sentences that belong to a particular class are equally similar to other sentences that belong to that class and sentences that do not. (b) Agreement prediction accuracy on reduced object RCs and unreduced object RCs as a function of $\mathbb{D}(RC, \neg RC)$

entirely by the presence of these function words.

In order to test whether the similarity between members of classes was indeed being driven by the presence of shared function words, we compared the representation space of the models we tested in the previous sections (henceforth *trained models*) with the representation space of models trained on no data (henceforth *baseline models*). Since the baseline models were only ever exposed to the 20 sentences in the adaptation set and there was no lexical overlap in content words between adaptation and test sets, any similarity between sentences in the representation space of these models would be driven by the presence of function words. If the similarity between sentences in the representation space of the trained models was being driven by factors other than the presence of function words, we would expect this similarity to be greater than the similarity between these sentences in the representation space of the baseline models.

We cannot directly use adaptation effect to compare the similarity between sentences in the representation spaces of trained models and baseline models, however: models trained on more data are likely to have stronger priors and are therefore less likely to drastically change their representations after 20 sentences than models trained on less data. In order to mitigate this issue, we defined a distance measure between sentences that belong to a class and sentences that do not belong to a class S_X as follows (see Figure 4.4 for a schematic):

$$\mathbb{D}(S_X, \neg S_X) = \frac{AE(X_2 \mid X_1)}{AE(\neg X_2 \mid X_1)}$$

This value would be greater than one if sentences that belonged to a class were more similar to each other than they were to sentences that did not belong to the class. Since the strength of prior belief would affect sentences that belong to the class the same way it would affect sentences that do not belong to the class, the effect would

cancel out.

We measured the distance between members and non-members for three linguistically interpretable classes: sentences which contained the same type of RC, sentences that matched in their reduction or sentences that contained any type of RC. In our baseline models, for all three classes, sentences that belonged to one of these classes were more similar to each other than sentences that did not belong to that class (Figure 4.5a). This was surprising for the class of sentences that contained any type of RC because there was no function word that was shared by all sentences in this class. We hypothesize that this is because sentences without RCs always contained the word *and*, whereas sentences with RCs never did.

In cases where members of the class shared at least some function words, the distance between sentences that belonged to the class and sentences that did not for the trained models was greater than that for the baseline models. This suggests that the similarity between sentences in the representation space of trained models was being driven by factors other than the mere presence of function words. However, somewhat surprisingly, as the number of training tokens increased, the distance between members and non-members decreased.

In the case where the members of the class did not share any function words, the distance between sentences that belonged to the class and sentences that did not belong to the class did not differ between the trained models and the baseline models. This suggests that any similarity between sentences in the representation space of trained models was driven purely by the presence (or in this case absence) of lexical items.

4.5.6 Does $\mathbb{D}(\boldsymbol{RC}, \neg\boldsymbol{RC})$ predict agreement prediction accuracy?

Marvin and Linzen (2018) created a dataset that evaluated the grammaticality of the predictions of language models. Using this dataset, they showed that LSTM LMs could not accurately predict the number of the main verb if the main clause subject was modified by an object RCs (either reduced or unreduced). However, the models had better performance if the main clause was modified by an active subject RC. For example, the models were at near chance levels in predicting that (15a) should have higher probability than (15b), but were slightly better at predicting that (16a) should have higher probability than (16b):

(15) a. The farmer that the parents love swims.

b. *The farmer that the parents love swim.

(16) a. The farmer that loves the parents swims.

b. *The farmer that loves the parents swim.

One possible explanation for this poor performance is that object RCs, either reduced or unreduced, are quite infrequent Roland, Dick, and Elman (2007). If the LM treats object RCs as unrelated to other RCs, there are likely very few training examples from which the models can learn about subject-verb agreement when the subject is modified by an object RC. If the LM had instead treated object RCs as belonging to the same class as other RCs, it could learn to generalize from training examples of subject-verb agreement when the subject is modified by other RCs. This suggests the hypothesis that agreement prediction accuracy on object RCs will be higher in LMs in which the representation of object RCs is more similar to the

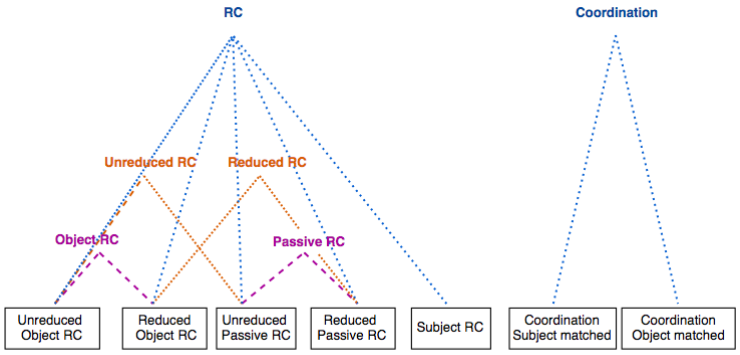


Figure 4.6: A schematic of how sentences belonging to different linguistically defined classes are related to each other in the LMs' representation space. Each colour indicates a different level of hierarchy.

representation of other RCs.

The similarity between object RCs and other RCs was defined as in the previous section (the proportion of blue squares to pink squares of the top two rows in Figure 4.4). There was an increase in accuracy as the number of hidden units increased (see Figure 4.5b). However, the similarity between object RCs and other types of RCs did not significantly correlate with agreement prediction; we therefore did not find any evidence for the hypothesis mentioned above.[10]

4.6 Discussion

Drawing on the syntactic priming paradigm from psycholinguistics, we proposed a new technique to analyze how the representations of sentences in neural language models (LMs) are organized. Applying this paradigm to sentences with relative clauses (RCs), we found that the representations of these sentences were organized in

[10]Similar patterns were observed for the other constructions in the dataset. See Supplementary Materials.

a linguistically interpretable hierarchical manner (summarized in Figure 4.6).

We investigated whether this hierarchical organization was driven by function words that are shared among sentences sentences or whether there was evidence that LMs were tracking more abstract properties of the sentence. We found that for at least some linguistically interpretable classes, sentences that belonged to these classes were more similar to each other in the representation space of the LMs we tested than in the representation space of baseline LMs that were not trained on any data. This suggests that the trained LMs were capable of tracking abstract properties of the sentence.

However, for linguistically interpretable classes in which sentences shared a non-lexically observable property (e.g. presence of a gap), sentences were as similar to each other in the representation space of the LMs we tested as in the representation space of baseline LMs. Taken together, these results suggest that LMs might be able to track abstract properties of classes of sentences only if these classes also share a lexically observable property.

Additionally, we found that the sentences belonging to linguistically interpretable classes were more similar to each other in the representation spaces of models trained on 2 million tokens than in the representation spaces for models trained on 20 million tokens. We infer from this that LMs' ability to track abstract properties of sentences decreases with an increase in the training corpus size. This suggests that if we want these LMs to track more abstract linguistic properties, training them on more data from the same distribution is unlikely to help (cf. Schijndel, Mueller, and Linzen (2019)). Future work can explore how to bias these models to track linguistically useful properties through architectural biases Dyer et al. (2016), training on auxiliary tasks Enguehard, Goldberg, and Linzen (2017) or data augmentation Perez and Wang

(2017).

We hypothesized that models' accuracy on subject verb agreement when preceded by object RCs would increase as the similarity between object RCs and the other types of RCs increased. However, we did not find evidence for this. This could either be because the similarity between object RCs and the other types of RCs was too weak to be useful (see Figure 4.5a) or because the LMs do not use this property when predicting verb agreement. Future work can disambiguate these reasons by testing models that are biased to treat sentences with object RCs and other RCs as being similar.

Finally, our method allows us to generate a similarity matrix in the LMs representation space for any given set of structures. In the future, generating a similar matrix for human representations using priming experiments and comparing these two matrices using analysis methods from cognitive neuroscience Kriegeskorte, Mur, and Bandettini (2008) may enable us to gain insight into how human-like the LM representations are and vice versa.

4.7 Conclusion

We proposed a novel technique to analyze how the representations of various syntactic structures are organized in neural language models. As a case study, we applied this technique to gain insight into the representations of sentences with relative clauses in RNN language models and found that the representations of sentences were organized in a linguistically interpretable manner.

4.8 Acknowledgments

We would like to thank Sadhwi Srinivas and the members of the CAP lab at JHU for helpful discussions and valuable feedback.

4.9 Appendix

4.9.1 Templates

We created seven templates (one for each of the structures we tested) to generate the adaptation and test sets. Each template had seven slots: subject, object of the relative clause, object of the main clause, verb in the relative clause, verb in the main clause, adverb for the main clause and adverb for the relative clause. The adverb arguments were blank strings half the time. The seven templates varied in the order in which they combined these arguments together to form a sentence. Therefore, for a given set of arguments, we were able to generate seven lexically matched sentences with different structures.

We included several sources of noise in our sentence generation process.

- Each noun slot was filled by a plural noun 40% of the time.
- Every noun phrase was modified with an adjective with 50% probability and every adjective was further modified with an intensifier with 40% probability.
- In cases when a verb (in the main clause or relative clause) was modified by an adverb, the adverb occurred pre-verbally or post-verbally with equal probability.

The slots in the templates were filled by 223 verbs, 164 nouns, 24 adverbs and 78 adjectives. In order to ensure semantic plausibility, we created sub-classes of nouns, adverbs and adjectives and manually specified which sub-classes could combine together. For example, the noun subclass “human" consisted of the nouns *friend*, *cousin*, *partner*, *sibling* and *colleague*. This class could serve as subjects for 38 verbs and could be modified by four sub-classes of adjectives. Similarly the verb

congratulated could take the noun subclass “human" as its subject and the noun subclasses “scienceperson" and “power" and as its object (e.g., *scientist*, *researcher* etc.; *principal*, *manager* etc.). Additionally, it could be modified by adverb subclasses “sad" and “time" (e.g, *sadly*,*gloomily* etc.; *yesterday*, *last month* etc.)

We ensured that there was no lexical overlap between adaptation and test sets, apart from function words (like *the*, *and*, *by*, *that* etc) and intensifiers (like *very*, *rather*, *quite etc*). We also ensured that verbs, nouns, adverbs and adjectives were not repeated within the same sentence.

4.9.2 Relationship between $\mathbb{A}(Y \mid X)$ and *Surp*(Y) prior to adaptation

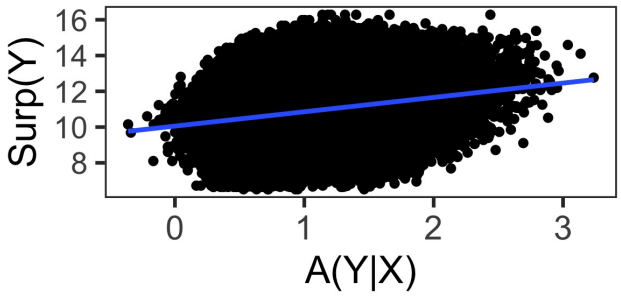


Figure 4.7

LM formula:

$A(Y \mid X) \sim center(Surp(Y))$

Results:

$\hat{\beta} = 0.061, SE = 0.0003, p < 2e-16$

4.9.3 Statistical Analyses:

This section contains details about the statistical analyses for all the results described in the main paper. In describing the formula for our mixed effects models we use standard LMER notation.

4.9.3.1 Validating AE as a similarity metric

For this analyses we fit a separate LMEM for each of the different structures that models could get adapted to.

LMER formula:

AE $\sim$ structure + (1 | adaptlist) + (1 | clist)

- Structure is a categorical variable coded as 1 if the test structure is the same as the adaptation structure and -1 if it is different.
- adaptlist: Which of the 10 adaptation-test sets we generated was the model adapted to and tested on?
- clist: Which subset of Wikipedia was the model trained on?

Structure adapted to	$\hat{\beta_{structure}}$	*SE*	p-value
Unreduced Object RC	0.256	0.001	$p < 2e-16$
Reduced Object RC	0.171	0.001	$p < 2e-16$
Unreduced Passive RC	0.229	0.001	$p < 2e-16$
Reduced Passive RC	0.100	0.001	$p < 2e-16$
Active Subject RC	0.194	0.001	$p < 2e-16$
Subject coordination	0.147	0.001	$p < 2e-16$
Object coordination	0.145	0.001	$p < 2e-16$

4.9.3.2 Similarity between sentences with different types of VP coordination

We fit the following mixed effect model on LMs that were adapted to sentences with coordination.

LMER formula:

AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable coded as 1 if the model was tested on sentences with RCs and -1 if the model was tested on sentences with the other type of coordination (e.g, for model adapted to ASRC-matched coordination, testtype was -1 if it was tested on PS/ORC-matched coordination)

$\hat{\beta} = -0.173, SE = 0.0007, p < 2e-16$

4.9.3.3 Similarity between sentences with different types of RCs

We fit the following mixed effect model on LMs that were adapted to sentences with RCs.

LMER formula:

AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable coded as 1 if the model was tested on sentences with other types RCs (e.g., for a model adapted to unreduced object RC, the value of testttype was 1 when tested on reduced object RC, reduced/unreduced passive RC and active subject RC). It was coded as -1 if the model was tested on sentences with coordination.

$\hat{\beta} = 0.038, SE = 0.0004, p < 2e-16$

4.9.3.4 Similarity between sentences belonging to different sub-classes of RCs

We fit the following mixed effect model on LMs that were adapted to sentences with object or passive RCs.

LMER formula:

AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable with four levels: passive match, reduced match, no match and both match. Since there were four levels, there were three contrasts. Passive match was chosen as the baseline and coded as 0 for all of the contrasts. For each contrast, one of the other levels was coded as 1 — i.e. in each contrast, the mean adaptation effect of passive match was compared to the mean adaptation effect of one

of the other conditions.

Contrast	$\hat{\beta_{testtype}}$	*SE*	p-value
Reduced match	-0.058	0.001	$p < 2e-16$
Both match	0.171	0.001	$p < 2e-16$
No match	-0.143	0.001	$p < 2e-16$

Table 4.2: Analysis 5.4

4.9.3.5 What properties of sentences drive the similarity between them?

We a separate mixed effects model for each of the three linguistically interpretable classes discussed in Section 5.5 of the paper. We did not include the baseline models in these analyses.

LMER formula:

$\mathbb{D}(S, \neg S) \sim$ scale(nhid) * scale(csize) + (1 | adaptlist) + (1 | clist)

nhid refers to the number of hidden units (100, 200, 400, 800, 1600) and csize refers to the training corpus size in millions of tokens (2, 10, 20).

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	0.008	0.002	$p = 0.003$
csize	-0.011	0.001	$p = 0.00002$
nhid:csize	-0.012	0.001	$p = 0.00001$

Table 4.3: $\mathbb{D}(RC, \neg RC)$

4.9.3.6 Does $\mathbb{D}(RC, \neg RC)$ predict agreement prediction accuracy?

We fit a separate linear regression model for LMs adapted to either reduced or unreduced Object RCs.

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	0.016	0.001	$p < 2e-16$
csize	-0.006	0.001	$p = 0.00004$
nhid:csize	-0.008	0.001	$p < 0.00001$

Table 4.4: $\mathbb{D}(\textit{Reduced match}, \neg\textit{Reduced match})$

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	-0.007	0.002	$p = 0.008$
csize	-0.023	0.002	$p < 2e-16$
nhid:csize	-0.040	0.001	$p < 2e-16$

Table 4.5: $\mathbb{D}(RC_X, RC \neq X)$

LM formula:

accuracy $\sim \mathbb{D}(RC, \neg RC)$ + scale(nhid) + scale(csize)

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}(RC, \neg RC)$	-0.007	0.098	$p = 0.947$
nhid	0.057	0.007	$p \ll 0.0000001$
csize	0.001	0.008	$p = 0.879$

Table 4.6: Models adapted to unreduced object RCs

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}(RC, \neg RC)$	-0.084	0.113	$p = 0.465$
nhid	0.013	0.005	$p = 0.018$
csize	-0.004	0.008	$p = 0.489$

Table 4.7: Models adapted to reduced object RCs

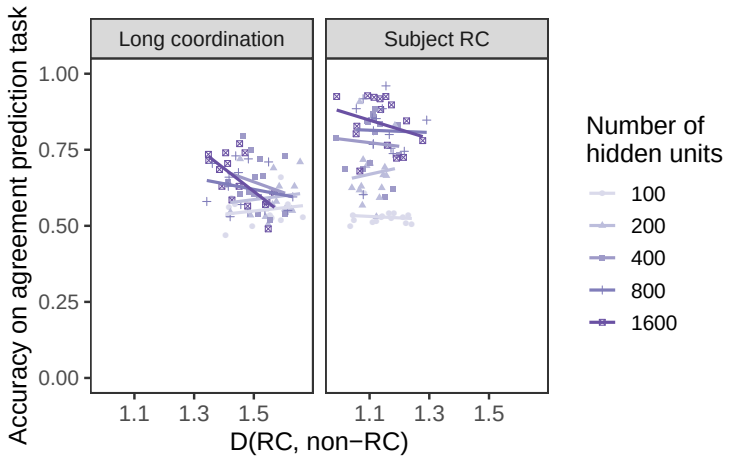


Figure 4.8

4.9.4 Relationship between $\mathbb{D}(RC, \neg RC)$ and agreement prediction accuracy for other structures

LM formula:

accuracy $\sim \mathbb{D}(RC, \neg RC)$ + scale(nhid) + scale(csize)

Predictor	$\hat{\beta}_{testtype}$	SE	p-value
$\mathbb{D}(RC, \neg RC)$	-0.215	0.204	$p = 0.297$
nhid	0.089	0.013	$p \ll 0.0000001$
csize	0.016	0.013	$p = 0.211$

Table 4.8: Models adapted to unreduced active subject RCs

Predictor	$\hat{\beta}_{testtype}$	SE	p-value
$\mathbb{D}(RC, \neg RC)$	-0.125	0.110	$p = 0.259$
nhid	0.023	0.008	$p = 0.014$
csize	0.025	0.008	$p = 0.003$

Table 4.9: Models adapted to unreduced sentences with long coordination

References

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema (2018). "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 240–248. URL: https://www.aclweb.org/anthology/W18-5426.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018). "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. URL: https://www.aclweb.org/anthology/P18-1198.

Shi, Xing, Inkit Padhi, and Kevin Knight (2016). "Does String-Based Neural MT Learn Source Syntax?" In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1526–1534. DOI: 10.18653/v1/D16-1159. URL: https://www.aclweb.org/anthology/D16-1159.

Kádár, Ákos, Grzegorz Chrupała, and Afra Alishahi (2017). "Representation of Linguistic Form and Function in Recurrent Neural Networks". In: *Computational Linguistics* 43.4, pp. 761–780. DOI: 10.1162/COLI_a_00300. URL: https://www.aclweb.org/anthology/J17-4003.

Qian, Peng, Xipeng Qiu, and Xuanjing Huang (2016). "Analyzing Linguistic Knowledge in Sequential Model of Sentence". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 826–835. DOI: 10.18653/v1/D16-1079. URL: https://www.aclweb.org/anthology/D16-1079.

Lakretz, Yair, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni (2019). "The emergence of number and syntax units

in LSTM language models". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 11–20. URL: https://www.aclweb.org/anthology/N19-1002.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies". In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535. DOI: 10.1162/tacl_a_00115. URL: https://www.aclweb.org/anthology/Q16-1037.

Weber, Noah, Leena Shekhar, and Niranjan Balasubramanian (2018). "The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models". In: *Proceedings of the Workshop on Generalization in the Age of Deep Learning*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 24–27. DOI: 10.18653/v1/W18-1004. URL: https://www.aclweb.org/anthology/W18-1004.

McCoy, R. Thomas, Robert Frank, and Tal Linzen (2018). "Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks". In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Ed. by Tim Rogers, Marina Rau, Jerry Zhu, and Chuck Kalish. Austin, TX, pp. 2093–2098.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. DOI: 10.1162/tacl_a_00254. URL: https://www.aclweb.org/anthology/Q19-1004.

Alishahi, Afra, Grzegorz Chrupała, and Tal Linzen (2019). "Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop". In: *Journal of Natural Language Engineering* 25.4, pp. 543–557.

Bock, J. Kathryn (1986). "Syntactic persistence in language production". In: *Cognitive Psychology* 18.3, pp. 355–387.

Wells, Justine B., Morten H. Christiansen, David S. Race, Daniel J. Acheson, and Maryellen C. MacDonald (2009). "Experience and sentence processing: Statistical learning and relative clause comprehension". In: *Cognitive Psychology* 58, pp. 250–271.

Kaschak, Michael P., Timothy J. Kutta, and John L. Jones (2011). "Structural priming as implicit learning: Cumulative priming effects and individual differences". In: *Psychonomic Bulletin & Review* 18.6, pp. 1133–1139.

Schijndel, Marten van and Tal Linzen (2018). "A Neural Model of Adaptation in Reading". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4704–4710. URL: https://www.aclweb.org/anthology/D18-1499.

Branigan, Holly P. and Martin J. Pickering (2017). "An experimental approach to linguistic representation". In: *Behavioral and Brain Sciences* 40.

Marvin, Rebecca and Tal Linzen (2018). "Targeted Syntactic Evaluation of Language Models". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1192–1202. URL: https://www.aclweb.org/anthology/D18-1151.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018). "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1195–1205. DOI: 10.18653/v1/N18-1108. URL: https://www.aclweb.org/anthology/N18-1108.

Jumelet, Jaap and Dieuwke Hupkes (2018). "Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 222–231. URL: https://www.aclweb.org/anthology/W18-5424.

Wilcox, Ethan, Roger Levy, Takashi Morita, and Richard Futrell (2018). "What do RNN Language Models Learn about Filler–Gap Dependencies?" In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 211–221. URL: https://www.aclweb.org/anthology/W18-5423.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy (2019). "Neural language models as psycholinguistic subjects: Representations of syntactic state". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 32–42. URL: https://www.aclweb.org/anthology/N19-1004.

Mahowald, Kyle, Ariel James, Richard Futrell, and Edward Gibson (2016). "A meta-analysis of syntactic priming in language production". In: *Journal of Memory and Language* 91, pp. 5–27.

Tooley, Kristen M and Matthew J Traxler (2010). "Syntactic priming effects in comprehension: A critical review". In: *Language and Linguistics Compass* 4.10, pp. 925–937.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*. Pittsburgh, PA: Association for Computational Linguistics, pp. 1–8. URL: https://www.aclweb.org/anthology/N01-1021.

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106, pp. 1126–1177.

Chang, Franklin, Gary S. Dell, and Kathryn Bock (2006). "Becoming syntactic". In: *Psychological Review* 113.2, p. 234.

Grave, Edouard, Armand Joulin, and Nicolas Usunier (2017). "Improving Neural Language Models with a Continuous Cache". In: *Proceedings of the Fifth International Conference on Learning Representations*. Ed. by Yoshua Bengio and Yann LeCun. International Conference on Learning Representations. URL: https://openreview.net/pdf?id=B184E5qee.

Krause, Ben, Emmanuel Kahembwe, Iain Murray, and Steve Renals (2017). *Dynamic Evaluation of Neural Sequence Models*. Tech. rep. University of Edinburgh. URL: https://arxiv.org/pdf/1709.07432.pdf.

Chowdhury, Shammur Absar and Roberto Zamparelli (2019). "An LSTM Adaptation Study of (Un)grammaticality". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 204–212. URL: https://www.aclweb.org/anthology/W19-4821.

Schijndel, Marten van, Aaron Mueller, and Tal Linzen (2019). "Quantity doesn't buy quality syntax with neural language models". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics.

Pinheiro, José, Douglas Bates, et al. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media.

Roland, Douglas, Fredric Dick, and Jeffrey L. Elman (2007). "Frequency of basic English grammatical structures: A corpus analysis". In: *Journal of Memory and Language* 57.3, pp. 348–379.

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith (2016). "Recurrent Neural Network Grammars". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 199–209. DOI: 10.18653/v1/N16-1024. URL: https://www.aclweb.org/anthology/N16-1024.

Enguehard, Émile, Yoav Goldberg, and Tal Linzen (2017). "Exploring the Syntactic Abilities of RNNs with Multi-task Learning". In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 3–14. DOI: 10.18653/v1/K17-1003. URL: https://www.aclweb.org/anthology/K17-1003.

Perez, Luis and Jason Wang (2017). "The effectiveness of data augmentation in image classification using deep learning". In: *arXiv preprint arXiv:1712.04621*.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A. Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* 2, p. 4.

Chapter 5

Conclusion and future work

At the broadest level, the goal of this dissertation is to further our understanding of human sentence comprehension. In the introduction, I argued that in order to understand *how* humans comprehend sentences, it is important to characterize *what* incremental structures humans implicitly build when processing sequences of words over time and *why* we build these structures. Then, I presented three projects in Chapters 2,3, and 4 that tackled these *what* and *why* questions by studying incremental structure building in sentences with different types of relative clauses. In this final chapter, I summarize the conclusions from the three projects and propose future work.

5.1 Summary of results

The *what* question

In Chapter 2, we tackled the question of what structures comprehenders construct by proposing a method for generating and testing hypotheses about the grammar that best describes the incremental structures that human comprehenders build. This method can be broken down into three steps: first, implement the representational

hypotheses from theoretical syntactic theories in an explicit model of parsing; second, generate *quantitative* behavioural predictions from this model; and finally, test these behavioural predictions by collecting empirical human sentence processing data.

We applied this method to study the incremental structures that comprehenders build when processing sentences with reduced relative clauses, such as (1).

(1) The graduate student examined by the committee proposed a framework.

We started by introducing two competing hypotheses in theoretical syntax about the underlying structure of sentences like (1): the Whiz-Deletion and the Participle-Phrase hypotheses. Then, we implemented the representational assumptions of these two accounts in a new model of parsing that we developed — SPAWN — in which parsing decisions are influenced by computational principles within the general cognitive architecture of ACT-R. Unlike the existing model of ACT-R based parsing proposed by Lewis and Vasishth (2005), SPAWN includes an explicit mechanism for processing null elements, a feature that is necessary for implementing the representational assumptions of the two accounts we considered. Next, we generated behavioural priming predictions from the Whiz-Deletion and Participle-Phrase versions of the model — specifically, the extent to which comprehenders are expected to parse an ambiguous sequence like "the graduate student examined" as having a reduced RC reading (i.e., the graduate student was being examined) as opposed to a main verb reading (i.e., the graduate student was doing the examining), when this ambiguous prompt was preceded by different types of prime sentences. Finally, we tested these predictions in a large-scale experiment ($N = 765$) using a novel web-based comprehension-to-production paradigm.

The qualitative behavioural predictions from the Whiz-Deletion account, but not the Participle-Phrase account, aligned perfectly with the empirical data from our experiment. Given that the predictions from the theoretical accounts were generated from an interpretable computational model in which the link between model behaviour and the hypothesized representations and processes is transparent, we can gain insights beyond merely stating that the Whiz-Deletion account better describes the structures that human comprehenders build than the Participle-Phrase account. Below we describe two such additional insights that we gained from our results.

First, by describing why the model predicts the qualitiative pattern of results it does, to the extent that this pattern aligns with human behavioural patterns, we gained insight into what factors might influence the way people process or produce sentences in experimental settings. For example, in both the Whiz-Deletion version of our model and in the human behavioural data, the priming effect was highest in passive reduced RCs, followed by an equal priming effect in passive full RCs and progressive reduced RCs, with the lowest priming effect in sentences without RCs. In our model, this priming pattern was a consequence of two distinct assumptions: the difference between sentences without RCs and sentences with RCs was driven by the representational assumption that all RCs share a common structure (i.e., a CP node), which is absent from sentences without RCs; on the other hand, the difference between passive reduced RCs and the other two types of RCs was a consequence of the statistics of these structures in the training data and the processing mechanism, and not a direct consequence of the representational assumption as such. The fact that the predicted priming pattern was a result of the interaction between models' grammar and the specific parsing mechanism we assumed suggests that a similar

interaction might be driving the empirical priming results with human comprehenders. This challenges the assertion made by Branigan and Pickering (2017) that behavioural priming results can provide "an implicit method of studying linguistic representation" even when there is no explicit model of parsing that implements these representational assumptions.

Second, by comparing the *quantitative* and not just the *qualitative* alignment between the model's predictions and the empirical data, we gained insight into the consequences of making the simplifying assumptions we did about our parsing mechanism. While the Whiz-Deletion version of our model perfectly predicted the qualitative pattern of priming, it underestimated the *magnitude* of the difference in priming between sentences with and without relative clauses. One likely cause for this difference is our simplifying assumption that parsing is strictly serial. Under this assumption, the probability of a reduced RC parse is determined entirely by the probability that the model assigns the category NP/CP to the subject noun (e.g., *the graduate student* in the target sequence "the graduate student examined"). However, as we discuss in § 2.5.3, given prior psycholinguistic evidence, we expect that the parsing decision should be influenced by the probability of the model assigning the category TP/VoiceP to the ambiguous verb (i.e., *examined*). In order for the parsing decision to be influenced by the categories assigned to the noun and the verb, the model needs to main multiple parses, and therefore goes against our strictly serial parsing assumption. This suggests that in order to fully account for the empirical results, it is necessary to implement a parallel parsing mechanism, which is also a suggestion made by previous work Boston et al. (2011).

The *why* question

This dissertation also indirectly explored two different types of questions about why comprehenders construct the structures they do. First, why might the incremental structures that comprehenders build change depending on the comprehenders' environment? Second, why do comprehenders use the grammar that they do when constructing incremental structures?

Studying the effect of comprehenders' environment on incremental structure building. In the first chapter of this dissertation, I discussed one possible explanation for why the incremental structures that comprehenders construct can change depending on the environment they are in: under a rational account of sentence comprehension, the optimal strategy for a parser processing an ambiguous sequence is to construct an incremental structure consistent with the highest possible parse more often than other structures consistent with the lower probability parses.[1] Since the probability of parses can vary drastically across environments and contexts, it is necessary for rational comprehenders to use context-specific probability distributions when building incremental structures.

If the above explanation for the influence of the environment on comprehenders' parsing decisions is accurate, then we would expect to find the following empirical pattern: participants who are repeatedly exposed to reduced relative clauses in an experiment will learn to expect more reduced relative clauses in the experimental setting and consequently construct incremental structures consistent with the reduced relative clause parse more often than participants who were exposed to filler sentences

[1] Or under a parallel parsing account construct all or many of the possible structures but weight according to the probability of the parses they are consistent with.

without relative clauses. The evidence for this empirical pattern is controversial. Early work by Fine et al. (2013) used the self-paced reading paradigm and found the predicted difference between the two groups of participants. However, later work by Stack, James, and Watson (2018), which used the same paradigm but included more items and participants, failed to replicate this effect. There are two possible reasons for this failed replication. First comprehenders do not rapidly update their expectations, and the between-group difference that Fine et al. (2013) found was a Type-I error; this explanation, if true, challenges the rational account. Second, comprehenders do update their expectations, but this update results in very small changes to their behaviour in the self-paced reading paradigm, making this effect difficult to detect even with the larger number of participants in Stack, James, and Watson's experiment.

The goal of Chapter 3 was to clarify this empirical picture. In a large self-paced reading we found evidence that participants exposed to sentences with reduced relative clauses did indeed assign reduced relative parses to temporarily ambiguous sequences more often than participants exposed to filler sentences. This evidence supports the prediction that participants rapidly update their expectations to match the statistics of their environment (*expectation adaptation*), thereby supporting the rational account of sentence comprehension. We also demonstrated in this chapter that self-paced reading was not an ideal paradigm to study this expectation adaptation because the change in reading times due to expectation adaptation was confounded with the change in reading times due to task adaptation (i.e., increased familiarity with the experimental paradigm); power simulations indicated that future experiments designed to detect modulations of the basic expectation adaptation effect could be underpowered with even 1200 participants. In light of this observation, the novel

comprehension-to-production paradigm proposed in Chapter 2 might be better suited to study this expectation because, unlike in self-paced reading, in this paradigm expectation adaptation is not confounded with task adaptation.[2]

Studying why comprehenders use the grammar they do In Chapter 2 we inferred that the grammar that shapes the incremental structures that comprehenders construct when processing sentences with reduced RCs is more consistent with the representational assumptions of the Whiz-Deletion account than those of the Participle-Phrase account. In Chapter 1 I described two not-mutually-exclusive hypotheses for why this might be the case: first, the grammar that comprehenders use is shaped by memory limitations during language acquisition; and second, the grammar is shaped by the statistics of the linguistic data participants have been exposed to. The goal of Chapter 4 was to propose a method for testing such hypotheses using neural network models trained to predict upcoming words.

Modern neural network models have been very successful on a variety of natural language understanding tasks, even achieving "superhuman" performance on many evaluation datasets (Wang et al., 2019). Although more stringent and targeted evaluations have revealed that these models are far from perfect (Kim and Linzen, 2020; McCoy, Pavlick, and Linzen, 2019; Marvin and Linzen, 2018; Warstadt et al., 2020), their strong natural language learning capabilities, especially when compared to their symbolic counterparts, makes them ideal objects of study on which we can run experiments that we otherwise cannot run on humans (cf. "animal models"; McCloskey

[2]If anything, task adaptation in this paradigm predicts the *opposite* pattern of behaviour change compared to the change predicted by expectation adaptation: as participants got used to the experimental paradigm, they might produce fewer completions consistent with the reduced RC parse because they might discover that the easiest way to complete the task is to generate completions with one word, which are consistent with the main verb parse.

1991). For instance, one way of testing the two hypotheses described above is to systematically alter the models' memory capabilities and linguistic input and measure the subsequent change in the incremental structures that these models construct.

In order to use these models in the manner described above, it is necessary for us to be able to study the incremental structures that these models construct in the first place, which is not trivial because the representations of these models are very large matrices that are not easy to interpret. There has been a lot of work in the recent years focused on making these matrices more interpretable (for review see Belinkov and Glass 2019; Rogers, Kovaleva, and Rumshisky 2020). Building on this work, we proposed a new method to characterize the incremental structures that these models construct when processing sentences, which is inspired by the priming paradigm from psycholinguistics. In this method we measured the probability that the model assigns to words in target sentences before and after the model was trained on a small set of prime sentences. If the probability for words in the target sentences increases after being trained on the prime sentences, then we can infer that something that the model learned from the prime sentences was useful when processing the target sentences. By ensuring that there is no lexical overlap between the prime and target sentences, we can infer from an increase in target probability post-priming that being exposed to the *structure* of the prime sentences made the *structure* of the target sentences more predictable.

We applied this method to different neural network models to study how the predictability of words in target sentences with relative clauses changed when they were trained on prime sentences with different structures. We found that the predictability of words in sentences with any specific type relative clause (e.g., passive reduced

relative clause) increased the most when models were trained on the same type of relative clause. Crucially, consistent with our empirical priming results from Chapter 2, we found that the predictability of words in sentences with reduced relative clauses increased more when they were trained on full relative relative clauses, than when they were trained on minimally different sentences without relative clauses. This suggests that the grammar that describes the structures constructed by the neural network models we tested, like the grammar that describes the structures constructed by human comprehenders, better aligned with the representational assumptions of the Whiz-Deletion account. In future work, I plan to study *why* this is the case by systematically altering the neural networks' memory capabilities and linguistic input and measuring change in priming behaviour.

5.2 Future work: moving beyond relative clauses

Apart from the future work proposed above, two other directions for future work present themselves from the methods and results in this dissertation. The first direction involves testing the extent to which the parsing mechanism and the Whiz-Deletion version of the grammar we assumed in Chapter 2 can account for other empirical data involving reduced relative clauses. For example, recent work demonstrated that the garden path effect in sentences with reduced relative clauses — i.e., the increased reading times in the temporarily ambiguous reduced relative clauses compared to unambiguous full relative clauses — was much larger than the garden path effect in other temporarily ambiguous sentences (Huang et al., 2022). Related work demonstrated that neural network parallel processing models without any re-analysis mechanism were able to predict the *direction* of garden path effects across the different types of

temporarily ambiguous sentences, but not the *magnitude* of this difference, leading the authors to conclude that a re-analysis mechanism was likely required to fully explain garden path effects (Van Schijndel and Linzen, 2021). Since the SPAWN model we proposed in Chapter 2 has an explicit re-analysis mechanism, future work can evaluate whether this model can predict both the direction and magnitude of the difference in garden path effects. It is very unlikely that the model in its current for can capture the magnitude of these effects because of the simplifying assumptions we made in our implementation of the parsing mechanism and the data we trained our model on. Nevertheless, systematically testing what changes need to be made to the model in order to fully capture the empirical garden path data, can shed light on what properties are crucial for characterizing how human comprehenders process temporarily ambiguous sentences.

The second direction involves building on the grammar we assumed to account for a wider range of psycholinguistic phenomena. There are at least two possible approaches extending the existing grammar. The first approach involves starting with existing grammars used in Natural Language Processing tasks (e.g., the grammar used in CCGBank; Hockenmaier and Steedman 2007), evaluating what psycholinguistic phenomena SPAWN models implemented with these grammars can and cannot capture, and use these results to isolate specific parts of the grammar that need to be improved. The second approach involves identifying other phenomena with competing theoretical accounts, and evaluating them. This approach can contribute additional data points to the efforts of identifying which of the theoretical debates in syntactic theory are relevant for sentence processing and which are not (cf., Graf, Monette, and Zhang 2017).

5.3 Conclusion

This dissertation includes three projects which study incremental structure building during sentence comprehension using a variety of computational approaches and experimental paradigms. While the focus of the phenomena studied in these projects was very narrow — they all studied structure building in sentences with reduced relative clauses — I hope that the methods proposed in this work can be used in future work to study incremental structure building more broadly.

References

Lewis, Richard L and Shravan Vasishth (2005). "An activation-based model of sentence processing as skilled memory retrieval". In: *Cognitive science*, pp. 375–419.

Branigan, Holly P and Martin J Pickering (2017). "An experimental approach to linguistic representation". In: *Behavioral and Brain Sciences* 40.

Boston, Marisa Ferrara, John T Hale, Shravan Vasishth, and Reinhold Kliegl (2011). "Parallel processing and sentence comprehension difficulty". In: *Language and Cognitive Processes* 26.3, pp. 301–349.

Fine, Alex B., T. Florian Jaeger, Thomas A. Farmer, and Ting Qian (2013). "Rapid Expectation Adaptation during Syntactic Comprehension". In: *PLoS One* 8.10, e77661. URL: https://doi.org/10.1371/journal.pone.0077661.

Stack, Caoimhe M. Harrington, Ariel N. James, and Duane G. Watson (2018). "A failure to replicate rapid syntactic adaptation in comprehension". In: *Memory and Cognition* 46.6. DOI: 10.3758/s13421-018-0808-6.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2019). "Superglue: A stickier benchmark for general-purpose language understanding systems". In: *Advances in neural information processing systems* 32.

Kim, Najoung and Tal Linzen (2020). "COGS: A Compositional Generalization Challenge Based on Semantic Interpretation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9087–9105. DOI: 10.18653/v1/2020.emnlp-main.731. URL: https://aclanthology.org/2020.emnlp-main.731.

McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448. DOI:

10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334.

Marvin, Rebecca and Tal Linzen (2018). "Targeted Syntactic Evaluation of Language Models". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1192–1202. DOI: 10.18653/v1/D18-1151. URL: https://aclanthology.org/D18-1151.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: A Benchmark of Linguistic Minimal Pairs for English". In: *Proceedings of the Society for Computation in Linguistics 2020*. New York, New York: Association for Computational Linguistics, pp. 409–410. URL: https://aclanthology.org/2020.scil-1.47.

McCloskey, Michael (1991). "Networks and theories: The place of connectionism in cognitive science". In: *Psychological science* 2.6, pp. 387–395.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. DOI: 10.1162/tacl_a_00254. URL: https://www.aclweb.org/anthology/Q19-1004.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A primer in bertology: What we know about how bert works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866.

Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen (2022). "SPR mega-benchmark shows surprisal tracks construction- but not item-level difficulty". In: *The 35th Annual Conference on Human Sentence Processing*.

Van Schijndel, Marten and Tal Linzen (2021). "Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty". In: *Cognitive Science* 45.6, e12988.

Hockenmaier, Julia and Mark Steedman (2007). "CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank". In: *Computational Linguistics* 33.3, pp. 355–396. DOI: 10.1162/coli.2007.33.3.355. URL: https://aclanthology.org/J07-3004.

Graf, Thomas, James Monette, and Chong Zhang (2017). "Relative clauses as a benchmark for Minimalist parsing". In: *Journal of Language Modelling* 5.1, pp. 57–106.