ABSTRACT

Title of Dissertation:	CONTEXT-DRIVEN EXPECTATIONS IN REAL-TIME SENTENCE PROCESSING: INVESTIGATIONS OF ADULTS, CHILDREN, AND LARGE LANGUAGE MODELS
	Eun Kyoung Lee Doctor of Philosophy, 2025
Dissertation directed by:	Professor Colin Phillips Department of Linguistics

A central feature of human language processing is the ability to rapidly build and update complex representations based on prior context. While humans actively use various linguistic cues to guide expectations, they sometimes fail to show immediate sensitivity to contextual information. A striking example is the role-reversal phenomenon, where comprehenders appear to expect verbs that fit the opposite argument roles indicated by the preceding context. These errors have often been attributed to failures in context-based representation building processes.

This dissertation investigates how context-driven expectations unfold in real-time, focusing on the mechanisms that support anticipatory processing and where vulnerabilities arise. Using the role-reversal phenomenon as a test case, I unpack the generation and integration processes into distinct stages, in order to determine which

parts are robust and which are more prone to errors. Through a combination of behavioral, electrophysiological, developmental, and computational methods, I examine divergences in role-sensitivity between comprehension and production measures, between adults and children, and between humans and computational models of language processing. The findings overall demonstrate that context-driven representations are generally robust and that errors tend to arise in later stages—particularly when new input must be integrated into prior representations under conditions of weak expectations. Moreover, the variability across different measures, sentence types, and populations—including children and large language models—can be traced to differences in the strength and timing of these underlying processes.

Together, the results indicate that predictive failures do not reflect a general weakness in the representation-building system but rather stem from specific points of vulnerability. These insights help to refine our understanding of how context-driven expectations operate in real-time and contribute to broader theories of the cognitive mechanisms that support human language processing.

CONTEXT-DRIVEN EXPECTATIONS IN REAL-TIME SENTENCE PROCESSING: INVESTIGATIONS OF ADULTS, CHILDREN, AND LARGE LANGUAGE MODELS

by

Eun Kyoung Lee

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Colin Phillips, Chair
Associate Professor Naomi Feldman
Associate Professor Ellen Lau
Professor Philip Resnik
Associate Professor L. Robert Slevc, Dean's Representative

Acknowledgements

I had thought pursuing a PhD would be going on a long solo journey. I turned out to be wrong. So many great people held my hand and supported me in so many different ways that I will forever be grateful for.

First, I would like to thank Colin Phillips, my advisor and committee chair, for the immense support and kindness he has given me throughout the past seven years of my master's and doctoral studies at UMD. Colin is a great teacher—I was so inspired by the discussion questions he asked us in Psycholinguistics I and II that I decided I really wanted to try to answer those questions and that he was the one who could help me do it. I learned from him how to think critically, write powerfully, mentor students (e.g., bringing snacks to lab meetings makes people happy) and importantly, that creating a strong community is essential to good science. I cherished the time we spent thinking through difficult questions together, even while taking walks around campus in the rain, brainstorming new experimental paradigms, and writing and reshaping our work to present to different audiences. It was a lot of fun working with Colin, and I feel truly grateful to have had an advisor who I always felt had my back.

Thank you to Naomi Feldman, Ellen Lau, Philip Resnik, and Bob Slevc, my committee members, for their support, insightful feedback, and encouragement throughout the different stages of this work. I always came away from meetings with Naomi with a clearer sense of direction—her questions and suggestions helped me further develop my ideas and focus on what truly mattered. I learned how to think like a computational psycholinguist, and I always felt that my writing, slides, and posters improved dramatically with her feedback. Ellen showed me what it means to truly love science and to question the basic assumptions in the field that I had taken for granted. From her, I learned how to look at the bigger picture without sacrificing attention to the small but crucial details. Philip motivated me to think more deeply about the connection between human and machine language processing. I would love to create a setting similar to his seminar in the future, where people with different perspectives from computational and human science backgrounds come together to discuss common questions. I enjoyed my conversations with Bob and his perspectives on production made me think about my work in new ways, which I appreciated.

Thank you to the faculty and mentors in the UMD Department of Linguistics and the UMD Department of Second Language Acquisition who have provided me with such generous support and a wonderful environment for graduate study and research. I am deeply grateful to Professors Mike Long, Nan Jiang, Robert DeKeyser, Kira Gor, and Steve Ross for their support during my master's studies. Thank you to Howard Lasnik, Norbert Hornstein, Jeff Lidz, Alexander Williams, Valentine Hacquard, Bill Idsardi, Masha Polinsky, Andrea Zukowski, Tonia Bleam, Kate Mooney, Aron Hirsh, and Omar Agha for their support throughout my doctoral studies. Thank you to Peggy Antonisse for being such a caring teaching advisor. And a big thank you to Kim Kwok for always being there to help, whether it was coordinating travel for conferences, processing funding requests, or even just getting the office printer working.

Thank you to my cohort—Imane Bou-Saboun, Xinchi Yu, Luisa Seguin, Clara Cuonzo, Jack Ying, and Leslie Famularo. It was so encouraging to go through graduate school with such a diverse and supportive group of friends, and the fact that we were all international students added an extra layer of fun to our conversations in so many ways.

Thank you to my fellow psycholinguists and collaborators in the department, including Masato Nakamura, Hanna Muller, Tal Ness, Katherine Howitt, London Dixon, Sathvik Nair, Utku Turk, Allison Dods, Sebastian Mancha, Carmen Tang, and Novak Shi, many of whom were involved in parts of this dissertation work. Having such a wonderful group of collaborators made the projects not only more productive, but genuinely exciting and fun. I could write a whole chapter about how grateful I am for the memorable moments we shared in our weekly meetings—the intellectual discussions, the laughter, and the playful jokes that made the work so much more enjoyable. Also, thank you to Xiaoyu Yang, Joselyn Rodriguez, and Jingyi Chen for the moments we shared, whether it was sharing beds at conferences, exploring new venues, or simply chatting about life during department events. Those memories made this journey even more meaningful.

I am also grateful for the support from the UMD Language Science Center and the broader language science community. Thank you to Shevaun Lewis and Caitlin Eaves for ensuring that everything ran smoothly during the many LSC events that helped bring our community together.

Thank you to all my collaborators and friends outside UMD. This includes Charlotte Vaughn, Hannah Mechtenberg, Jessy Contreras, Stella Huang, and the students who participated in the summer museum course in 2022 and 2023, assisting with data collection at the Planet Word museum. Thank you to Sol Lago and Elise Oltrogge at the University of Frankfurt for the fun and intellectually engaging collaboration. I am also grateful for the friends I made at the University of Oxford during the 2024 summer research internship program. Thank you for welcoming me and Allison so warmly and making our time there such an enjoyable and memorable experience.

Playing tennis was the most effective way to take a break during graduate school, and that was only possible because of the amazing friends I had in the UMD Korean Graduate Student Tennis Club and the Jjoleb group. I am grateful for the sense of belonging this group brought into my life.

Thank you also to Minsun Kim, Atousa Motameni, Parvaneh Oliazadeh, Joanne Sung Eun Kim, and all my friends and family in South Korea for their love, support, and encouragement.

Thank you to my mentors from my undegraduate studies at Seoul National University, Professors Sungeun Lee, Sun-Young Oh, and Youngsoon So, for their guidance in the early stages of my academic journey. I also thank the Kwanjeong Educational Foundation for the financial support throughout my studies.

I feel incredibly lucky to have met Chan Kim—my partner, my best friend, and my tennis buddy—during graduate school. The love and support he gave me kept me motivated and made me feel like I could achieve so many things I might have doubted otherwise. He genuinely cared about the questions that kept me up at night in my research, and his feedback helped me think through my work more carefully. It was a joy to go through the PhD journey together, and I am so proud of us.

Finally, I would like to thank my parents, Inki Lee and Sun-Young Lee, and my little brother, Eun Jun Lee. I cannot put into words how thankful I am for my supportive and loving family. I truly could not have done this without their support. Thank you for standing by me through all these years, and thank you, Mom and Dad, for helping me become who I am, giving me the courage to believe I can achieve anything in the world, and believing in me no matter what.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A hallmark of human cognition is the ability to process information rapidly, efficiently, and with remarkable accuracy. Humans not only construct complex mental representations in real-time but also dynamically update them as new information becomes available. In language processing, decades of research have shown that comprehenders do not simply react to incoming words—they actively predict what will come next, using prior context to shape expectations. Similarly, speakers often plan multiple words ahead before articulating an utterance. This ability to generate anticipatory representations is crucial for both speaking and understanding. Despite its efficiency, research has documented cases where speakers and listeners fail to optimally use available information in constructing anticipatory representations, leading to systematic errors during language processing. Why do humans sometimes fail to use readily available cues to constrain expectations? And what aspects of the underlying mechanism make it prone to such systematic failures? This dissertation seeks to answer these questions by examining a particularly striking case of processing failure: the role-reversal phenomenon. By analyzing this test case, I aim to uncover the cognitive mechanisms that drive the efficiency of real-time sentence processing, while also identifying its points of vulnerability.

In this chapter, I first introduce some foundational perspectives on context-driven incremental processes in sentence processing. Specifically, I highlight two key features of the underlying processes: i) the parser's tendency to actively incorporate new information and act upon it as soon as it becomes available, even at the risk of premature commitments, and ii) the process of adjusting initial interpretations when new input requires substantial revisions to previously constructed representations. Building on this background, I discuss how the role-

reversal phenomenon presents a unique challenge to our previous understanding of how these processes unfold in real-time. I argue that identifying the key steps involved in generating and integrating new information as a sentence unfolds is essential for pinpointing where systematic errors, such as role-reversals, originate. In the final section of this chapter, I propose a staged architecture that outlines the sub-processes that are assumed to operate in sentence processing. Using this framework, I categorize existing theoretical accounts of role-reversal errors based on which step they attribute role-reversal errors to. In the end, I propose which stages of the process are more robust and which are more vulnerable to errors, setting the stage for the empirical investigations in the subsequent chapters of this dissertation.

1.1 Context-driven expectations

Prediction is a term generally used to refer to the act of anticipating upcoming inputs before they appear. Research shows that listeners and readers actively predict what might come next, using prior context to guide their expectations. Studies using eye-tracking (e.g., Altmann & Kamide, 1999) and electrophysiological measures (e.g., Kutas & Hillyard, 1984) provide robust evidence for this anticipatory behavior. For instance, Altmann and Kamide (1999) demonstrated that listeners shift their gaze toward a likely referent before the corresponding noun is heard, indicating that they predict upcoming content based on verb semantics. Similarly, EEG studies have shown that unexpected words elicit a greater ERP response called the N400 (Kutas & Hillyard, 1984; DeLong et al., 2005), suggesting that comprehenders continuously generate and update expectations as they process sentences. These findings indicate that prediction is an integral part of sentence comprehension, allowing individuals to construct linguistic representations efficiently in real-time.

A large body of empirical evidence further highlights how proactive the parser is in constructing anticipatory representations. Comprehenders sometimes commit to certain structures prematurely such that it sometimes leads to misinterpretations that require later revision, which is not always successful, and such errors can result in lingering processing difficulties. One well-studied phenomenon that illustrates this is the garden-path effect, where comprehenders initially adopt an incorrect syntactic structure and must later revise their interpretation. For example, in sentences like, *The horse raced past the barn fell* (Bever, 1970), readers often interpret *raced* as the main verb when it is first encountered rather than as part of a reduced relative clause (*The horse [that was] raced past the barn...*), only to later realize that their initial parse was incorrect. When confronted with the unexpected word *fell*, comprehenders must reanalyze the sentence structure, which often leads to processing slowdowns and lingering confusion (Frazier & Rayner, 1982; Christianson et al., 2001). This demonstrates the interplay between strong early commitments and the difficulty in processing new inputs that do not fit previous built representations.

Anticipatory processing can extend beyond the immediately upcoming word, as comprehenders may predict multiple aspects of an upcoming phrase before fully encountering it. One example comes from the processing of determiner-noun combinations, where comprehenders anticipate a specific determiner before even encountering the noun that follows. DeLong and colleagues (2005) found that when readers expected a noun like *airplane* following a context, they also highly expected its associated determiner form-specifically (*an airplane* rather than *a airplane*). Upon encountering a determiner and noun that did not match their expected determiner-noun pairing (e.g., *a kite* instead of *an airplane*), this led to increased N400 amplitudes early on, at the determiner, compared to when the input matched their expectations. More recent studies suggest that revisions are not always immediate or complete, and comprehenders may continue to

show residual effects of their initial predictions even after encountering the input (Nieuwland et al., 2018).

Studies have shown that expectation generation extends beyond minimal combinatorial structures, involving proactive construction of larger syntactic representations. As a result, people can expect specific types of elements based on the structure they are currently building. For example, in filler-gap dependencies, comprehenders anticipate certain types of verbs based on preceding syntactic cues. Omaki et al. (2015) showed that upon encountering a wh-phrase (e.g., *The book that the author...*), comprehenders expect a transitive verb, assuming that the sentence will continue with a direct-object structure (e.g., *The book that the author wrote…*), and are surprised when they encounter an intransitive verb (e.g., e.g., *The book that the author chatted about…*). This suggests that the parser does not merely wait for the verb to construct sentence representations but instead actively generates possible syntactic representations when there is information from the preceding context that makes a particular structure, and hence, a particular class of verbs, more likely than others.

The examples reviewed so far demonstrate the proactive approach that the parser takes in constructing representations moment-by-moment. They also show that mismatches between early-built representations and subsequent input lead to processing difficulty. Despite these risks, the parser actively constructs representations even in the presence of uncertainty during real-time language processing. The examples also highlight the parser's sensitivity to different types of contextual cues, which are used to generate expectations for upcoming input and to constrain expectations further or revise them when more information becomes available. Given this kind of general anticipatory behavior and the sensitivity to various contextual cues, it is surprising when we observe cases where comprehenders appear to fail to quickly constrain expectations in a

context-sensitive manner. One of the most striking cases of this is the role-reversal phenomenon, where comprehenders appear surprisingly blind to previously assigned thematic roles when generating expectations for upcoming verbs.

1.2 The role-reversal phenomenon: An apparent failure of context-driven processing

Thematic roles, such as agent and patient, provide crucial information about who is performing an action and who is affected by it. One would expect comprehenders to integrate these roles with world knowledge about event likelihoods to generate and refine predictions about unfolding sentences. Early work by Ferretti et al. (2001) demonstrated that comprehenders rapidly use thematic role information to facilitate verb prediction. Their findings showed that both event knowledge and verb argument structure shape expectations about likely agents and patients. Similarly, Kamide et al. (2003) used eye-tracking to show that listeners actively anticipate upcoming arguments based on verb meaning and syntactic constraints. However, subsequent research has revealed that this predictive process is not always optimal. In some cases, comprehenders appear insensitive to role-reversal anomalies, treating implausible thematic assignments and verbs as though they were expected.

One of the clearest pieces of evidence for the failure to constrain verb predictions using argument roles comes from EEG studies on the N400 response. The N400 ERP component, a neural marker of semantic processing difficulty, typically increases when a word is unexpected or incongruent with prior context (Kutas & Hillyard, 1984). However, several studies have found that role-reversed sentences, where agents and patients are swapped to describe unlikely events, fail to elicit the expected N400 effect, suggesting that comprehenders do not immediately detect thematic role anomalies. For example, Kuperberg et al. (2003) examined how the brain processes

conceptual relationships between noun phrases and verbs in simple, unambiguous sentences. Their study distinguished between anomalous sentences like 1a, thematic role animacy violations, where the NP was inanimate but assigned an agent role, and and 1b, non-thematic role pragmatic violations, where the NP was a plausible agent (i.e., boys can bury), but the verb was pragmatically incongruent with the prior context.

1a) For breakfast, the eggs would only eat...

1b) For breakfast, the boys would only bury...

While the non-thematic role pragmatic violations elicited a significant N400 effect, thematic role animacy violations instead triggered a significant P600 effect, typically associated with syntactic reanalysis and repair. The authors proposed that rather than immediately rejecting an implausible agent assignment, comprehenders attempted to reinterpret the sentence structure, potentially reassigning the noun's role from agent to theme. Similar findings were observed in later studies with different sentence constructions and languages (Hoeks et al., 2004; Kim & Osterhout, 2005; van Herten et al., 2005; Kuperberg, 2007; Brouwer et al., 2012; Chow et al., 2016, 2018), and with different experimental measures, such as anticipatory looks in the visual world paradigm (Kukona et al., 2011) or eye-tracking while reading (Burnsky, 2022).

Similar failures have also been observed in verb-final languages, where anticipating the verb may facilitate more efficient comprehension. In languages like German and Japanese, comprehenders must wait until the end of a sentence to encounter the verb, creating a potential incentive to generate expectations based on earlier cues. Yet, studies have shown that even when contextual information clearly signals who is acting upon whom, comprehenders can still exhibit temporary misanalysis when the verb appears with an unexpected argument structure (Nakamura et al., 2024; Stone & Rabovsky, 2025).

Another important characteristic of the role-reversal phenomenon is that the apparent blindness to argument role information occurs in early stages of prediction. This insensitivity does not persist; comprehenders do not generally maintain incorrect role assignments once the verb has been processed. Instead, the key issue concerns the timing of role-based expectations before the verb appears. Chow et al. (2018) provide evidence that comprehenders can use argument role information to constrain verb expectations, but only when sufficient time is available before verb onset. The authors tested native speakers of Mandarin using sentences like 2a and 2b, where thematic role assignment should strongly constrain expectations for the upcoming verb.

2a) 警察把小偷抓了回警局。

cop BA thief arrest (and bring back) to police station.

"The cop arrested the thief (and brought him back) to the station."

2b) 小偷把警察抓了回警局。

thief BA cop arrest (and bring back) to police station.

"The thief arrested the cop (and brought him back) to the station."

In sentence 2a, comprehenders should expect a verb describing an action performed by the agent (cop) on the patient (thief), whereas in sentence 2b, the verb should describe an event with thief as the agent and cop as the patient. ERP results showed no significant N400 differences when the same verbs were presented in both contexts, even though thematic roles made such predictions unlikely. However, the N400 effect was found between sentences that contained an additional prepositional phrase between the arguments and the verb (i.e., *thief BA cop ZAI that evening arrest*). The authors claimed that the additional material between the arguments and the verb created a time delay which allowed sufficient time for argument roles to constrain verb expectations, before the verb was encountered. Subsequent studies have demonstrated that simply introducing a time delay,

without adding any extra linguistic material, produces the same N400 effect (Liao et al., 2022; Nakamura et al., 2024; Stone & Rabovsky, 2025). This suggests that the time between the introduction of argument role information and verb onset plays a critical role in enabling immediate role-sensitivity.

The role-reversal phenomenon presents a unique challenge to dominant theories of predictive processing. The role-reversal phenomenon surprisingly contrasts with findings that highlight the parser's proactive approach to using contextual information in prediction. Given the strong evidence that comprehenders actively anticipate upcoming input based on syntactic structure, event knowledge, and semantic cues, one would expect them to use this information to constrain their verb predictions early on. However, the fact that comprehenders sometimes fail to immediately detect role-reversal anomalies suggests that some aspect of the predictive mechanism is more fragile than previously assumed. Understanding the precise locus of this failure is crucial for refining models of real-time sentence processing.

Before proceeding, one important clarification is necessary: the focus here and throughout the dissertation is on online processes rather than offline final interpretations. While the role-reversal phenomenon is commonly studied in real-time sentence comprehension, similar effects have also been observed in final sentence interpretations. For instance, studies on English passive sentences show that comprehenders incorrectly interpret the patient as the agent, despite clear morphosyntactic and lexical cues indicating otherwise. For example, in the sentence, *The chef was admired by the waiter*, despite passive morphology clearly marking *chef* as the patient, studies have found that comprehenders often misinterpret *chef* as the agent—as if the chef were doing the admiring rather than receiving admiration (Ferreira, 2003). This suggests that instead of fully integrating syntactic cues like passive morphology (*was admired by*), comprehenders may rely on

expectation-driven heuristics, such as assuming that the first noun in a sentence is the agent. While such offline role-reversal effects are well-documented in sentence interpretation (e.g., Meng & Bader, 2021), the primary focus of this dissertation is on online predictive failures—cases where comprehenders fail to constrain expectations when role information becomes available as a sentence unfolds in real-time.

1.3 Unpacking the underlying process

Different theories have been proposed to explain role-reversal effects in real-time sentence processing, each offering distinct explanations for why it occurs. Some accounts attribute these effects to faulty prediction mechanisms that over-rely on semantic associations rather than syntactic constraints, leading comprehenders to process verbs based on lexical co-occurrences rather than structural cues (e.g., Kim & Osterhout, 2005). Others suggest that comprehenders rely more on event knowledge than on linguistic structure, generating expectations based on typical real-world relationships rather than faithfully integrating syntactic cues (e.g., Kuperberg, 2007; Rabovsky et al., 2018; Stone & Rabovsky, 2025). Another perspective argues that noisy representations of the input lead to heuristic-based interpretations that diverge from the intended interpretations (e.g., Li & Ettinger, 2022). While later chapters provide a detailed discussion of specific theoretical accounts, here, I first introduce a framework that allows us to unpack the sub-processes involved in context-driven expectation generation and integration, which will help to pinpoint where role-reversal errors arise.

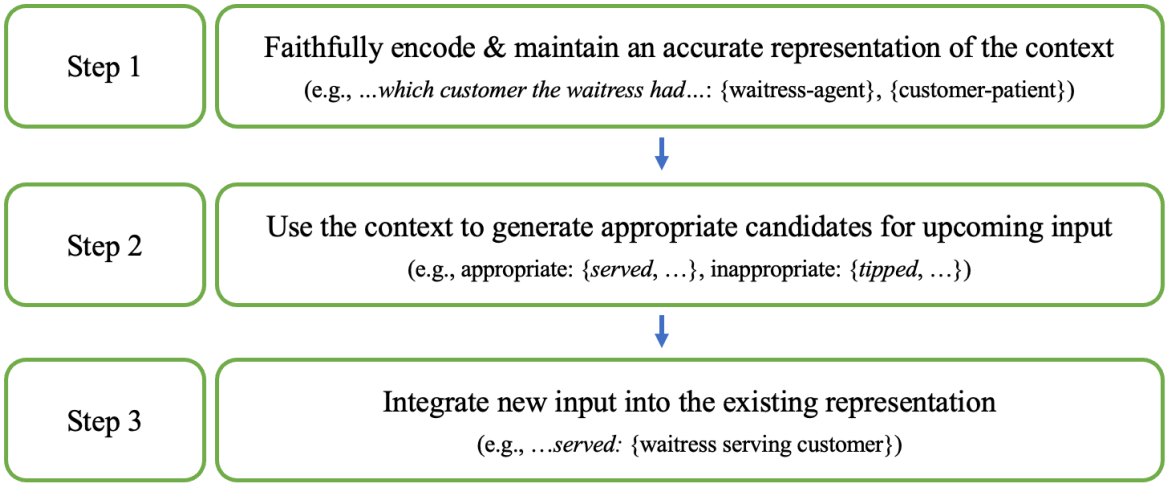


Figure 1: A schematic of the context-driven generation and integration process.

I identify three key steps required to construct and update representations in an anticipatory manner as a sentence unfolds (see Figure 1). **Step 1:** The first step is to accurately encode and maintain the preceding context. In the case of argument roles, this means assigning appropriate thematic roles to arguments (e.g., *waitress* = agent, *customer* = patient) and retaining that assignment throughout subsequent processing. **Step 2:** Using the representation from Step 1, the parser proactively generates expectations for upcoming input. These predictions should remain faithful to the contextual information. For instance, given a *waitress-agent* and *customer-patient* pair, verbs like *served* would be likely candidates, whereas role-inappropriate verbs like *tipped* should be less expected. **Step 3:** Once new input is encountered, the parser must integrate it into the existing representation, e.g., update the event structure to *"waitress serving customer"* upon encountering the verb *serve*.

Some accounts attribute role-reversal effects to specific stages in this pipeline. For example, Chow et al. (2016, 2018) and Nakamura et al. (2024) place the source of the error in Step 2, arguing that early verb predictions are not yet informed by argument role context and are instead shaped

by shallow lexical associations. Other accounts, such as the Sentence Gestalt model (Stone & Rabovsky, 2025), locate the error earlier in the pipeline, in Step 1, claiming that comprehenders form probabilistic expectations about likely thematic roles, which can lead to initial uncertainty or misassignments that persist into later processing. Similarly, Liao et al. (2022) argue that argument role representations are constructed too slowly to shape early predictions, leading to role-insensitive expectations at the time of the verb.

There are also accounts that are more difficult to assign to a single step in the pipeline. For example, Li and Ettinger (2021), adopting a noisy-channel framework, propose that comprehenders construct heuristic interpretations of sentences that may diverge from their literal structure. In these cases, role-reversal errors emerge not solely from faulty predictions or misrepresentations of the context, but from interpretive processes that integrate incoming input with noisy or degraded context representations. Likewise, some models emphasize the downstream misuse or reweighting of contextual cues in evaluating subsequent inputs. Kim and Osterhout (2005) suggest that semantic associations can override syntactic constraints, while Kuperberg et al. (2007) argue that alternative semantic-thematic mappings may compete with syntactic interpretations. Although these accounts are not strictly expectation-based and often describe interpretive outcomes rather than context-driven predictive processes, they converge on the idea that, even when context is correctly represented, it is not fully or effectively utilized to constrain interpretation at the verb.

One reason for the proliferation of theoretical accounts and the difficulty in adjudicating among them is that most empirical work has relied on a relatively narrow set of measures, often focused on responses to the verb itself. While much of the theorizing hinges on processes that unfold *before* the verb appears (e.g., how context is encoded, what verb candidates are considered),

the majority of the existing evidence comes from what happens *after* comprehenders are confronted with the verb. This has limited our ability to isolate the specific stages at which role-reversal errors originate. Moreover, although many models highlight the importance of world and event knowledge, empirical work has rarely examined item-wise differences to test these claims. As a result, there has been relatively few tools for teasing apart the fine-grained underlying processes.

This dissertation addresses these limitations by expanding the range of tasks and measures used to probe underlying processes, particularly examining processes that occur with the pre-verbal context, and by leveraging variability across items to better understand the underlying cause of role-reversal errors. By doing so, I aim to clarify *which* processes are most vulnerable, *when* they are vulnerable, and *why*. Specifically, I argue that Steps 1 and 2 are largely robust: the parser faithfully encodes, maintains, and utilizes preceding contextual cues to anticipate and interpret upcoming input. Instead, the primary source of vulnerability lies in Step 3—the stage at which new input is processed and integrated into the existing representation. More specifically, I propose that errors arise when this input is encountered before earlier steps have fully completed and before strong candidate representations have been generated. In the case of role-reversals, when sufficient contextual information is available and processed prior to the verb, the parser can generate expectations that align with the intended argument roles. However, errors are more likely when candidate generation is incomplete at the time the verb appears. This view aligns with non-expectation-based accounts that attribute errors to interpretive processes involving the full sentence, including the verb, rather than to deficiencies in pre-verbal processing (e.g., Kim & Osterhout, 2005; Li & Ettinger, 2021). While these accounts have raised this possibility, there has been a lack of clear empirical evidence to support it. Throughout this dissertation, I present

experimental and computational evidence demonstrating that context-driven representation building is more robust than previously assumed, and that the integration of new input is the critical point of vulnerability. These findings reinforce the reliability of Steps 1 and 2 and identify Step 3 as the primary locus of role-reversal errors.

1.4 Outline of the dissertation

As discussed in the previous section, while role-reversal effects have been widely documented, there is ongoing debate about where in the processing stream these errors originate. Do they stem from a misrepresentation of argument roles (Step 1)? Are they driven by difficulties during context-driven candidate generation (Step 2)? Or do they emerge only later, after the verb is encountered (Step 3)? This dissertation seeks to resolve these questions by systematically examining role-reversal errors across different experimental measures (N400 vs. cloze tasks), different populations (adults vs. children), and different processing systems (human vs. artificial intelligence).

Specifically, this dissertation examines the sources of variability observed across prior studies on role-reversal anomalies. Existing research has reported divergent findings depending on experimental measures, linguistic contexts, and participant populations. Investigating these variability patterns is crucial, as they provide key insights into the underlying mechanisms of predictive processing. The following chapters introduce the empirical designs and results that address these key questions and demonstrate how they support the proposed areas of robustness and vulnerability.

Chapters 2 and 3 investigate why different experimental measures produce apparently contrasting findings on role-reversal sensitivity. N400 studies have consistently reported delayed sensitivity to role-reversal effects, suggesting that comprehenders do not immediately use thematic

role information to constrain their expectations. In contrast, speeded cloze studies have shown that speakers can rapidly incorporate argument role information to predict upcoming words in real-time. This apparent contrast raises fundamental questions about linking hypotheses, as well as the underlying mechanisms responsible for these differences between comprehension and production measures. To resolve this puzzle, these chapters examine why role-sensitivity appears to vary across different experimental paradigms, particularly regarding the nature of the tasks and linguistic materials used across studies. By disentangling these factors, the empirical investigations provide insight into how and when comprehenders successfully use argument roles to constrain expectations.

Chapters 4 and 5 address the parallels between adults and children in predictive processing. While previous research has demonstrated that children, like adults, engage in anticipatory language processing, a key question remains: Do children rely on the same mechanisms as adults when generating expectations? Prior studies suggest that children's processing of argument role information is relatively robust, but it is unclear whether they use this information in the same way as adults to constrain predictions. These chapters investigate this question using speeded cloze tasks to examine children's sensitivity to argument roles in generating expectations. Beyond addressing developmental differences, this investigation also contributes to the broader discussion of why a temporal delay has been shown to be crucial in bringing back N400 effects (Chow et al., 2016; Nakamura et al., 2024). The findings shed light on the conditions under which comprehenders—both children and adults—succeed or fail to use thematic roles predictively.

Chapter 6 explores whether large language models, which are trained to predict upcoming words based on probabilistic distributions, exhibit human-like patterns in generating next-word predictions. Prior studies have shown that probabilistic measures derived from large language

models correlate strongly with human responses, including N400 amplitudes and reading times. This has led to claims that human sentence processing operates similarly to these models, relying purely on probabilistic prediction of the next input, and that human responses closely track the context-based probabilities. I challenge this view by using role-reversals as a test case to assess whether data-driven probabilistic prediction alone can account for systematic patterns in human language processing. The findings suggest that while large language models demonstrate strong predictive power, they lack a key distinction that is present in human language processing—generating anticipatory representations and integrating new information with previously constructed representations. I argue that this distinction is critical for explaining systematic errors like role-reversals, and the results support the claim that human language processing cannot be reduced to simple data-driven next-word prediction.

In Chapter 7, I synthesize the findings from the previous chapters and explain how each piece supports the central claim of this dissertation: parsing and context-based representation building processes are largely robust, and errors arise at specific points of vulnerability, particularly in integrating new inputs into existing representations. The results provide evidence for systematic constraints in anticipatory processing, where failures such as role-reversal errors occur under certain conditions, rather than as a general limitation of the sentence processing system. This final chapter also discusses the broader theoretical implications of these findings and outlines directions for future research. By refining our understanding of how predictive mechanisms interact with structural constraints in real-time language processing, this dissertation contributes to ongoing debates about the nature of linguistic representation and access in sentence processing.

Each chapter in this dissertation is written to stand on its own, allowing readers to engage with individual chapters independently. As a result, some redundancy exists between the

introduction and later chapters, particularly where background information or theoretical framing is reiterated. Readers who have already familiarized themselves with the introduction may choose to skip or skim overlapping sections as needed. At the time of writing, a version of Chapter 2 is a manuscript under review: *Lee, E.-K. R., & Phillips, C. (forthcoming). Argument role sensitivity in real-time sentence processing: Evidence from a hybrid comprehension and production task.* Chapter 6 has been published as a conference proceedings paper: *Lee, E.-K. R., Nair, S., & Feldman, N. H. (2024). A psycholinguistic evaluation of language models' sensitivity to argument roles.* In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 3262–3274). Miami, FL: Association for Computational Linguistics.

Chapter 2: Generating candidates and making early commitments

2.1 Introduction

Prediction is considered as one of the key processes involved in real-time language comprehension (Altmann & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007, 2022; Hale, 2001; Levy, 2008; Van Petten & Luka, 2012; Hickok, 2012; Pickering & Garrod, 2013; Pickering & Gambi, 2018; Huettig, 2015; Kuperberg & Jaeger, 2016; Ryskin & Nieuwland, 2023). Studies have shown that comprehenders use preceding contextual information to form expectations about upcoming linguistic input before it is explicitly presented. A consensus view is that different types of contextual information serve as cues to help comprehenders generate multiple candidates that are consistent with the given linguistic input as well as with knowledge about what are likely events in the world (Staub et al., 2015; Kuperberg & Jaeger, 2016; Van Petten & Luka, 2012; Roland et al., 2012; Frisson et al., 2017; Ness & Meltzer-Asscher, 2021; Nakamura, 2023). It is generally agreed that these multiple candidates are activated quickly and accurately, in proportion to their grammaticality and probability in that context. Finally, it is widely assumed that the effects of candidate activations are transparently reflected in a variety of different measures, including anticipatory looks in eye-tracking, N400 responses in EEG, various reading time measures (e.g., as mediated via surprisal measures), and cloze responses.

The present work examines in detail a specific phenomenon where some evidence seems to challenge these general assumptions. The case of role-reversals, where inverting the roles of two arguments in a sentence context should yield different expectations for the upcoming verb, seems to suggest that predictions are not always constrained to contextually appropriate candidates and that there are inherent differences between the measures we use to probe underlying processes.

This test case turns out to be useful for validating each of the main elements of the consensus view, and for gaining insights into underlying mechanisms.

2.1.1 Initial failure to use argument roles in prediction during comprehension

A large body of work in the EEG psycholinguistics literature has shown that comprehenders do not show different N400 amplitudes to verbs in role-appropriate and role-reversed contexts, in which all elements are kept identical except the nominal arguments which are reversed (e.g., Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Nakamura et al., 2024). For example, the verb *served* is role-appropriate given the context in (1), while it is role-inappropriate given the context in (2), which contains the same lexical items but with reversed argument roles.

(1) *The restaurant owner forgot which customer the waitress had…*

(2) *The restaurant owner forgot which waitress the customer had….*

The lack of the N400 effect with argument role-reversals is surprising, given ample evidence of a high degree of N400 sensitivity to probable vs. improbable continuations (e.g., Kutas & Hillyard, 1980, 1984; Van Petten & Luka, 2012; Federmeier, 2007, 2022). This insensitivity to role information is taken to be restricted to the initial stages of processing, as people do show sensitivity in later measures, such as in the P600 or in sentence-final interpretations. If comprehenders use the given argument role information in the preceding context, and if they are able to effectively compute its consequences, they should show a more reduced N400 response to a verb in its role-appropriate context compared to when the verb and the given argument roles indicate an unlikely event, e.g., customers serving waitresses. This phenomenon of insensitivity to role reversals has also been observed in other measures, such as in eye-tracking with the visual world paradigm (Kukona et al., 2011) or eye-tracking while reading (Burnsky, 2022), and has led psycholinguists

to propose different theories on the underlying mechanisms involved in verb prediction and the use of argument role information in the generation of possible verb candidates.

Different accounts for role-reversal anomalies proposed in the literature place the blame for the initial role-insensitivity on different factors. We focus on two key factors that have been consistently alluded to in prior work. One family of accounts claims that comprehenders sometimes fail to make role-appropriate predictions because of insufficient time (Chow et al., 2016, 2018; Liao et al., 2022; Nakamura et al., 2024). According to this view, argument roles are not used to generate candidates at the point when role-sensitivity is measured. Comprehenders initially activate verbs that are lexically associated with the preceding arguments irrespective of whether they fit the given argument roles (i.e., a 'bag-of-arguments' prediction), and role information only later affects the prediction process, constraining expectations to role-appropriate verbs. This could be either because parsing argument roles is slow (Liao et al., 2022), or because probing memory and selecting verb candidates that fit the given argument roles takes time (Chow et al., 2016, 2018). The key supporting evidence for this view is the finding that having more time before the verb, following the arguments, gives rise to an N400 effect (Chow et al., 2018; Liao et al., 2022; Stone et al., 2024; Nakamura et al., 2024). For example, in Chow et al.'s (2018) EEG study with Mandarin Chinese, inserting an adverbial phrase (e.g. *zai shangxingqi*; 'last week') between the arguments and the critical verb led to significant N400 amplitude differences to target verbs in their role-appropriate and role-reversed contexts. Similar results have been found in German (Stone et al., 2024) and in Japanese, where simply inserting a presentation delay between an argument and the verb, without additional linguistic content, elicited a significant N400 effect (Nakamura et al., 2024). These findings suggest that time is the crucial factor that determines the presence of immediate role-sensitivity.

Another key factor that has been emphasized when explaining the role-reversal phenomenon is the relative strengths of different contextual cues. Some accounts claim that role-inappropriate verb candidates are generated due to strong influences from semantic cues or probabilistic world knowledge which prevent the structural role information from constraining expectations (Kim et al., 2015; Kuperberg, 2016; Li & Ettinger, 2022; Rabovsky et al., 2018). In other words, role-specific candidates are generated without delay, but they are overwhelmed by the strength of role-independent candidates. For example, a particular argument or a set of arguments might activate a highly probable event in long-term memory which involves those arguments, independent of the roles assigned by the context. This could yield role-inappropriate predictions that violate the constraints enforced by the structural (role) information in the preceding linguistic context. While computational simulations incorporating this factor have been shown to reproduce the observed empirical patterns (Robovsky et al., 2018; Li & Etteinger, 2022), the empirical support is not as clear as the timing evidence mentioned previously.

While there appear to be different dimensions that might contribute to initial role-insensitivity, i.e., the speed at which the comprehender can parse or use argument role information to generate predictions, or the availability of other contextual cues that might prevent role information from accurately constraining expectations, it is generally assumed that initial predictions are role-insensitive because these factors are at play in the underlying prediction generation processes. Set against this background, it is surprising that production studies using the speeded cloze paradigm have shown that argument roles have an immediate impact on generating next-word continuations.

2.1.2 Rapid use of argument roles in speeded cloze production

While the majority of the work examining sensitivity to argument roles during real-time processing comes from EEG studies using the N400 as the key measure, production studies using the speeded cloze paradigm (Staub et al., 2015) have shown that argument role information can be quickly used in generating verb predictions. In a speeded cloze paradigm, participants see a sentence context presented word-by-word and then produce a continuation within a time limit. Studies using this paradigm have shown that participants overwhelmingly produce role-appropriate verbs as their completions and rarely produce role-reversal errors, or verbs that do not fit the preceding argument roles (Chow et al., 2015; Nakamura, 2023; Nakamura et al., 2024). Recently, Nakamura et al. (2024) sharpened the evidence for a contrast, by using the same materials and comparable timing parameters in distinct comprehension and production studies, and still showing different sensitivity profiles. The authors proposed that the contrast can be captured within the same activation model by varying the strength of a process that filters out role-inappropriate candidates. They further proposed that this is a kind of monitoring process, analogous to well documented processes in production (Nakamura et al., 2024).

The apparent contrast in role-sensitivity is unexpected if both cases involve the same predictive processes and if the experimental measures equally reflect those processes. N400 effects generally arise when words differ in fit to context. This is widely taken as evidence that comprehenders are sensitive to degree of fit. Therefore, when words that obviously differ in their fit elicit matched N400s, it suggests that the words were erroneously considered as equally fitting given the different contexts in that moment. Similarly, the leading account of speeded cloze response rates and latencies claims that these measures reflect the relative speed of activation of competing candidates (Staub et al., 2015). If a candidate (and its competitors) is equally strong in two different contexts, it should show similar response rates or speed at which it gets produced as

the continuation of the sentence. The stark contrast observed between the prior studies, where the EEG comprehension experiments suggest role-insensitive verb expectations and speeded cloze production experiments show otherwise, suggests that some part of this account must be incorrect. Importantly, the accounts that are designed to explain role-insensitivity in comprehension do not straightforwardly accommodate the finding of sensitivity in production. Moreover, we cannot take for granted that the insensitivity observed in comprehension and the sensitivity observed in production are the product of the same processes.

2.1.3 Potential explanations for the comprehension-production contrast

One possible reason why a contrast is observed between comprehension and production with regard to role-sensitivity is that people engage in predictive processes to different degrees in comprehension and production experiments because of different experiment settings. Comprehension and production experiments differ in several aspects, and there is some evidence in the literature that suggests that those differences could contribute to the observed comprehension-production contrast found with role-reversals.

2.1.3.1 Different engagement in predictive processes because of different experiment materials

One key difference between comprehension and production experiments measuring role-sensitivity involves the properties of experiment materials. In comprehension studies, participants often see as many ungrammatical or implausible sentences as grammatical or plausible sentences, because comprehension experiments require eliciting and comparing participants' responses to those different types of sentences. Studies examining sensitivity to argument roles, in particular, include the presentation of many role-reversed sentences (e.g., *The restaurant owner forgot which waitress the customer had *served*). The presentation of role-reversed sentences is necessary, in

order to observe responses to inappropriate target verbs and to infer the underlying processes that lead to such response patterns. Conversely, in a production experiment, participants do not see role-reversed or other anomalous sentences, since they themselves generate responses and the plausibility of the sentence depends on the completions that they produce.

Previous studies have found that responses to linguistic anomalies can be affected by the proportion of predictable and unpredictable sentences in the experiment (Lau et al., 2013; Ness & Meltzer-Asscher, 2021), or by participants' own awareness of the proportion of plausible sentences in the experiment (Hammerly et al., 2019), or perceptions of how likely the speaker will make an error (Hanulíková et al., 2012). For example, Brothers and colleagues (2017) found that self-paced reading times for predictable and unpredictable sentence-final words were less divergent, indicating a weaker predictability effect, when a greater portion of the experiment sentences ended with unexpected words, relative to an experiment with a balanced number of predictable and unpredictable sentence-endings. Other studies have reported similar findings in word-level processing, where the proportion of related or predictable word pairs presented during an EEG experiment modulated N400 responses to predictable and unpredictable word pairs (Lau et al., 2013; Ness & Meltzer-Asscher, 2021). These findings suggest that people's predictive behavior can be affected by the experiment setting in terms of the kinds of stimuli that participants are presented with.

Given this background, one possible explanation for the lack of immediate role-sensitivity in comprehension, in contrast with production, is that the presence of role-reversed sentences makes argument roles less reliable cues in a comprehension experiment setting than in a production experiment setting. When many of the experimental sentences contain unexpected role-inappropriate verbs, participants might strategically choose to rely less on argument role

information for prediction, in order to reduce prediction error. In contrast, in a speeded cloze paradigm, it would be advantageous to use all available contextual information to produce responses quickly and accurately, as the chosen continuations carry no penalty.

One way to test whether the different experiment materials lead to the contrast is to expose participants to the same materials, including anomalous sentences, and to measure their behavior in comprehension and production. A hybrid interleaved trials design achieves this goal, by allowing us to measure role-sensitivity in comprehension and production while participants see the same materials during the experiment.

2.1.3.2 Different engagement in predictive processes because of different tasks

Another contrast between experiments that measure role-sensitivity in comprehension and production is the task given to the participants. Comprehension experiments often involve reading and judging the plausibility of presented sentences. On the other hand, production studies using the speeded cloze paradigm require participants to produce their own sentence continuations under time pressure. It is possible that the different task demands involved in comprehension and production experiments affect the way people use contextual information such as argument roles to generate expectations for upcoming words.

A body of psycholinguistic literature suggests that people may adopt different predictive strategies depending on the task, which affects predictability effects observed in comprehension measures like the N400 effect (Kuperberg & Jaeger, 2016; Xiang & Kuperberg, 2015; Brothers et al., 2017; Lau et al., 2013; Ness & Meltzer-Asscher, 2021; but also see Van Wonderen & Nieuwland, 2023). For example, in Brothers et al.'s (2017) EEG experiment, participants were instructed to actively predict sentence-final words given a discourse and to respond whether their predictions matched the actual sentence endings that were presented after a delay. The authors

found that this elicited a larger N400 effect between predictable and unpredictable words, relative to when participants simply read the same sentences and answered true-or-false comprehension questions in a portion of the trials. In this way, prior literature suggests that different task demands can contribute to observing different predictive processing behavior. More specifically, the typical instructions for a speeded cloze task resembles the kind of instructions that Brothers et al. (2017) have found to elicit stronger predictability effects in comprehension, where participants were asked to actively generate next-word predictions. It is possible that this kind of task affects the predictive process which results in increased sensitivity to argument roles for verb prediction in a speeded cloze paradigm than in an EEG experiment.

We tested the possibility that the comprehension-production divergence arises from task-related differences using the same interleaved trials that also served to present the same materials to participants. Given that participants do not know in advance whether a given trial is a comprehension or production trial, they would have to process the context for all trials in the same way, which would allow them to quickly produce a continuation at any random trial. We observed whether engaging in this task would yield similar role-sensitivity in comprehension and production.

2.1.3.3 Same engagement in predictive processes but different measures

A third possibility is that participants in fact engage in predictive processes in the same way in comprehension and production experiments, but that the experimental measures of role-sensitivity tap into different processes. It is possible that, regardless of stimulus-related or task-related differences in comprehension and production experiments, participants process sentence contexts in the same way and carry out the same predictive processes to generate the same kinds of expectations based on prior context. If measures like the N400 and speeded cloze responses equally reflect those underlying prediction generation processes (Staub et al., 2015), they should show

similar patterns in role-sensitivity. However, a contrast might arise because those measures tap into other additional processes that are not purely context-driven. N400 amplitudes are responses to a particular target input presented in a particular context, whereas cloze responses are generated by participants themselves solely based on the context, without the explicit presentation of a target input. Responding to a subsequent input following a context could invoke additional processes that do not occur when producing a continuation. Therefore, even when the context is processed and expectations are generated in the same way, responses to presented target verbs might look different from responses that are self-produced, and hence they might show different degrees of role-sensitivity in comprehension and production measures. If this is the case, we should not expect to observe the same degree of role-sensitivity in comprehension and production, even when participants are put in the same task setting and engage in the same predictive processes.

2.1.4 The present study

The current study aimed to better understand the comprehension-production contrast seen in prior studies via a hybrid design in which participants do not know on any given trial whether they will be asked to comprehend or produce a continuation. This paradigm allowed us to measure role-sensitivity in comprehension and in production, while controlling for potential differences that arise from comprehension-specific and production-specific experiment settings, including the materials and tasks used in the experiments.

Experimental sentences were presented under two conditions, canonical or reversed. Each pair of arguments appearing in the canonical condition had a predictable target verb continuation (e.g., *...which customer the waitress had served*), which was role-inappropriate in the reversed condition, where the arguments were reversed in order (e.g., *...which waitress the customer had served*). Sentences presented in the comprehension trials included target verbs, while the sentence

contexts presented in the speeded cloze production trials did not include the target verbs (i.e., sentences were truncated before the target verb).

Different measures were used to detect role-sensitivity in comprehension and production. Following prior work, we used N400 responses to target verbs as a source of evidence for immediate role-sensitivity in comprehension. A smaller N400 response to target verbs in the canonical condition than in the reversed condition would indicate immediate use of argument roles in constraining expectations. For the production trials, we used response rates and response times (RTs) as measures of role-sensitivity. Two types of analyses were conducted with response rates and RTs. First, we examined all the responses that were produced and categorized them into role-appropriate and role-inappropriate responses. If argument roles were effectively used in predicting the upcoming verb, the majority of responses were expected to be role-appropriate, with rare role-inappropriate responses, similar to previous speeded cloze findings.

In addition, we compared how often and how quickly the target verbs, which were presented and used to measure role-sensitivity in the comprehension trials, were produced in the (role-appropriate) canonical and (role-inappropriate) reversed conditions in the production trials. Comparing the cloze percentages of target verb responses in the canonical and reversed conditions would, however, not be sufficient to detect role-sensitivity. This is because interpreting cloze probability as a direct measure of the predictability of a word given the context, without considering the strengths of other candidates also generated by the context, is misleading (Staub et al., 2015).

Based on the findings from a series of speeded cloze experiments, Staub et al. (2015) proposed that producing a speeded cloze response is similar to a race among candidates that individually accumulate activation based on their strengths given the context. The first candidate

to reach an activation threshold wins the race and gets produced as the cloze response. Importantly, each candidate's strength is determined by its speed of activation, and the relative speed of candidates that participate in the race is what determines the win percentage of a candidate (i.e., its cloze probability). Then, according to the model, cloze probability does not directly reflect the fit between a word and a context, because how often a candidate wins depends on how fast the alternative candidates are in the race. If a race is very competitive (has overall strong competitors or a particularly strong competitor), it becomes less likely for a candidate to win with the same speed that allows it to win a race with weaker candidates. Since different contexts generate different candidates with different strengths, this means that a candidate with similar activation strengths in different contexts could produce different cloze probabilities, because those win percentages always depend on the relative speed of the other candidates in the race.

Therefore, in order to correctly interpret cloze outcomes of the target verbs produced in the canonical and reversed contexts, we must understand the relative strengths of candidates generated in the canonical and reversed contexts. To do so, we examined the cloze percentages and RTs of role-neutral verbs, i.e., verbs that were equally predictable in both role contexts (e.g., *...which customer/waitress the waitress/customer had* <u>*seen*</u>). The cloze patterns of role-neutral verbs would indicate the relative competitiveness of the race in the canonical and reversed contexts, which would allow us to better interpret the cloze outcomes of the target verbs produced in those contexts. Details of the analysis and interpretations of speeded cloze RTs are presented in the data analysis and results sections.

2.2 Experiment 2-1

The goal of Experiment 2-1 was to examine whether presenting anomalous, role-reversed sentences in a speeded cloze paradigm would yield an increase in role-reversal errors, i.e., yield

role-insensitivity in production similar to comprehension. If seeing numerous instances of role-reversal anomalies during the experiment leads to an increase in role-reversal responses, it would suggest that the initial role-insensitivity in comprehension is partly due to the task setting which makes argument roles a less reliable source of context for prediction. If the presentation of role-reversals does not cause a change in speeded cloze response patterns and does not elicit more role-reversal errors, it is unlikely that the exposure to role-reversed sentences is the main factor driving the comprehension-production contrast.

2.2.1 Method

2.2.1.1 Participants

Sixty-seven native English-speaking adults (24-58 years old, mean age = 39) recruited on Amazon Mechanical Turk participated in the experiment. Eighteen participants were excluded from analysis due to missing data or failure to follow instructions, leaving 49 participants for analysis.

2.2.1.2 Materials

We used the same stimuli that were used in a previous study that failed to find role-sensitivity in the N400 measure during comprehension (Chow et al., 2016). This included 60 pairs of sentences where each pair included canonical and reversed argument roles. These sentences were used in the comprehension trials. For the production trials, the same sentences were truncated prior to the verb and presented as sentence fragments to elicit a cloze response. An example set of critical items in each condition is presented in Table 1. The final word of each sentence in the production trials was presented in red font, which served as a prompt indicating that a response should be produced following that word.

Table 1: Example critical items

Trial type	Condition	Sentence (fragment)	Example cloze response
Production	Canonical	The restaurant owner forgot which customer the waitress had	*served*
	Reversed	The restaurant owner forgot which waitress the customer had	*tipped* (**served*)
Comprehension	Canonical	The restaurant owner forgot which customer the waitress had served last night.	N/A
	Reversed	The restaurant owner forgot which waitress the customer had served last night.	N/A

The critical items were divided into four presentation lists, where each list contained 120 items that were evenly divided across the following combinations: comprehension-canonical, comprehension-reversed, production-canonical, production-reversed. Each list contained an additional 120 filler items. Half of the filler items were identical to the control items used in Chow et al. (2016), where sentences either ended in highly predictable or unexpected words, which were found to elicit the typical N400 effect. The other half of the filler items were sentences with varying lengths and structures. The complete set of items was counterbalanced across the four lists and presented in random order. Participants never saw the same argument pair in the same condition in the same type of trial more than once, and none of the filler items occurred in the same trial type for a participant.

2.2.1.3 Procedure

A hybrid task was designed which interleaved speeded cloze production trials with comprehension trials, through which the target verbs were presented in either a role-appropriate or role-reversed context. A critical part of the design was that the comprehension and production trials were mixed in the same experiment and presented in random order such that at the start of each trial participants did not know whether the trial was a comprehension trial or a production trial. This encouraged

participants to be prepared to provide a cloze response at the start of every trial, since in any given trial, they would see a red prompt word and have to quickly generate a cloze response.

The experiment was administered online using PCIbex (Zehr & Schwarz, 2018). The presentation of each trial began with a fixation mark '+' presented for 1000 ms, followed by the context presented word-by-word in RSVP format with a 530 ms SOA (300 ms per word, 230 ms blank between words). For the comprehension trials, the presentation continued until it reached the end of the sentence, at which participants were given two seconds to respond whether the sentence they read was plausible or implausible by pressing corresponding keys on the keyboard. The screen moved onto the next trial after participants responded or reached the time limit. For the production trials, subsequent to the context presented in RSVP, the prompt word appeared in red font, followed by an underline, which indicated participants had to produce a cloze response. After a 3-second time limit, the screen proceeded to the next trial. Figure 2 presents an illustration of the interleaved trials design.

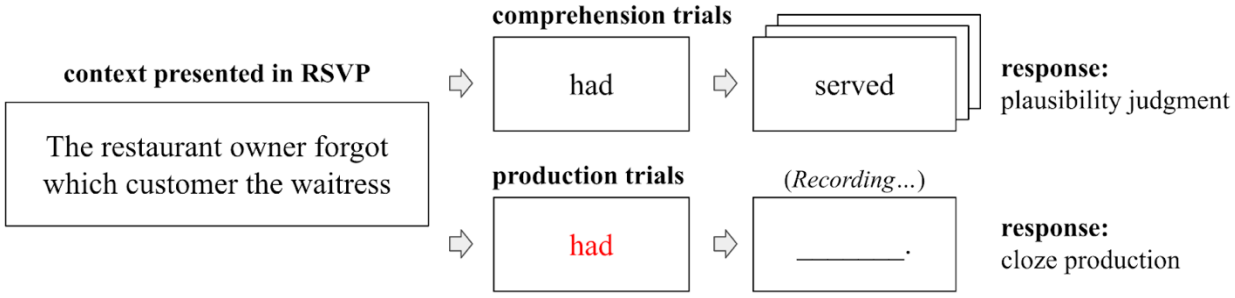


Figure 2: Illustration of the comprehension-production interleaved trials design in Experiment 2-1

Participants were familiarized with the task through a series of instructions and practice trials divided into multiple steps. The first part included instructions for carrying out a regular speeded cloze task, for the production trials of the experiment. Participants were told to read each sentence fragment presented word-by-word on the screen and then say aloud a likely continuation after they see a word in red. They were informed there was a time limit so that they should be fast

and accurate in their responses. The instructions were followed by sample recordings of responses and four subsequent practice trials. Participants were given the chance to adjust their microphone settings and recording quality prior to the experiment. Next, instructions for the comprehension trials were given, where participants were told that in some trials, they would see a complete sentence rather than the red word prompting a cloze response. For these trials, they were told to make plausibility judgments by pressing the 'J' (plausible) or 'F' (implausible) keys on the keyboard, where 'plausible' meant the sentence describes 'something that would happen normally.' Four practice trials were given for the comprehension trials. The final part of the instructions informed participants that they would see a mix of the production and comprehension trials throughout the experiment. They were told to be prepared to quickly provide a continuation of the sentence in every trial, in order to be able to produce a response before the time limit when they see a red word indicating that it is a production trial.

All participants provided written consent prior to the experiment and received monetary compensation for their participation at the end of the experiment. The complete experiment session took about one hour.

2.2.1.4 Analysis

Speech files containing cloze responses recorded in the production trials were pre-processed using Google Cloud Speech-to-Text API for automatic transcription and Chronset (Roux et al., 2017) for automatic detection of speech onset times. The pre-processed files were then manually checked and adjusted using the Praat software (Boersma, 2001). The complete speech processing pipeline is publicly available: https://github.com/ekrosalee/SpeechDataProcessingPipeline.

The cloze responses were coded and analyzed in the following ways. First, each unique response produced for a given pair of arguments was coded as i) biased toward the role assignment

in the canonical condition, ii) biased toward the role assignment in the reversed condition, or iii) not biased toward any particular role assignment. Then, all responses were coded as 'role-appropriate' when produced in the condition with the preferred argument role assignment, 'role-reversed' when produced in the opposite condition, or 'role-neutral' when it was a response that equally fit both role orders. All other responses were coded as 'other.'

In addition, the cloze percentages and RTs of 'target responses' were compared between the canonical and reversed conditions. Target responses were the same verbs that were presented in the comprehension trials, i.e., target verbs in Chow et al.'s (2016) EEG study, which did not elicit an N400 effect. Observing higher percentages of target verb responses in the canonical condition than in the reversed condition would not be sufficient evidence for role-sensitivity, because cloze percentages cannot be interpreted as direct reflections of the strength of candidates, independent of the other alternatives that were generated by the context (Staub et al., 2015). For example, it is possible that the target verbs were equally activated given the canonical context and the reversed context but showed different cloze outcomes because the competing candidates were different in the two conditions. Therefore, in order to better understand the candidate profiles in the canonical and reversed conditions, we examined the role-neutral responses produced in the two contexts. Role-neutral responses were verbs that did not have a particular bias for one context over the other (e.g., *…which customer/waitress the waitress/customer had* <u>*seen*</u>). The cloze percentages and RTs of role-neutral responses in the two conditions would provide a clearer picture of the kind of contexts in which the target verbs could get produced as a cloze response.

For all comparisons, statistical tests were carried out with the lme4 package (v1.1.35.1; Bates et al., 2015) in R (v4.3.2; R Core Team, 2023), using generalized linear mixed-effects models to compare response rates between conditions and linear mixed-effects models to compare

log-transformed RTs between conditions. All models initially included condition as a fixed effect, with a deviation coding (canonical = -.5, reversed = .5), and a maximally specified random effects structure with the subject and item random intercepts and by-subject and by-item slopes for condition. Random slopes were progressively dropped until the models converged (Barr et al., 2013). Significant effects were determined based on p-values ($p < .05$).

2.2.2 Results

In the comprehension trials, participants judged the sentences as plausible more in the canonical condition (89%) than in the reversed condition (35%) [$t(66) = 11.95$, $p < .001$)], indicating role-sensitive final interpretations. For the control items, participants judged the sentences as plausible more in the high-cloze condition (92%) than in the low-cloze condition (22%) [$t(66) = 13.86$, $p < .001$].

The production rates of the speeded cloze responses are plotted in Figure 3. A strong majority of responses were contextually appropriate given the preceding argument roles: role-appropriate and role-neutral responses together accounted for 88.5% of all responses, while role-reversed responses accounted for 11% which is similar to the 14.3% error rate found in a prior study using the same experimental materials without the interleaved production trials (Nakamura, 2023). Role-appropriate responses had the shortest onset latencies, while role-reversed and role-neutral responses had similarly longer RTs.

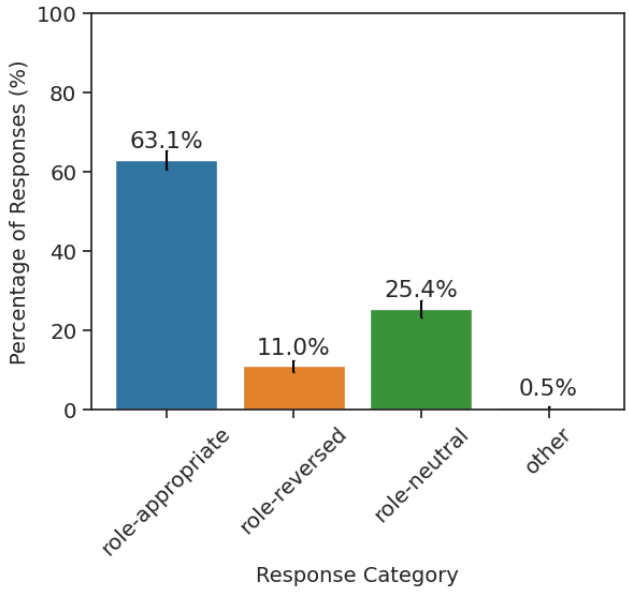


Figure 3: Percentages of cloze responses in each response category in Experiment 2-1. Error bars represent 95% confidence intervals.

Production rates of target responses diverged between the canonical and reversed conditions (Figure 4). Response rates showed a significant main effect of condition (β = -2.03, *SE* = .23, $p < .001$), where target responses were produced with a higher percentage in the canonical condition than in the reversed condition.

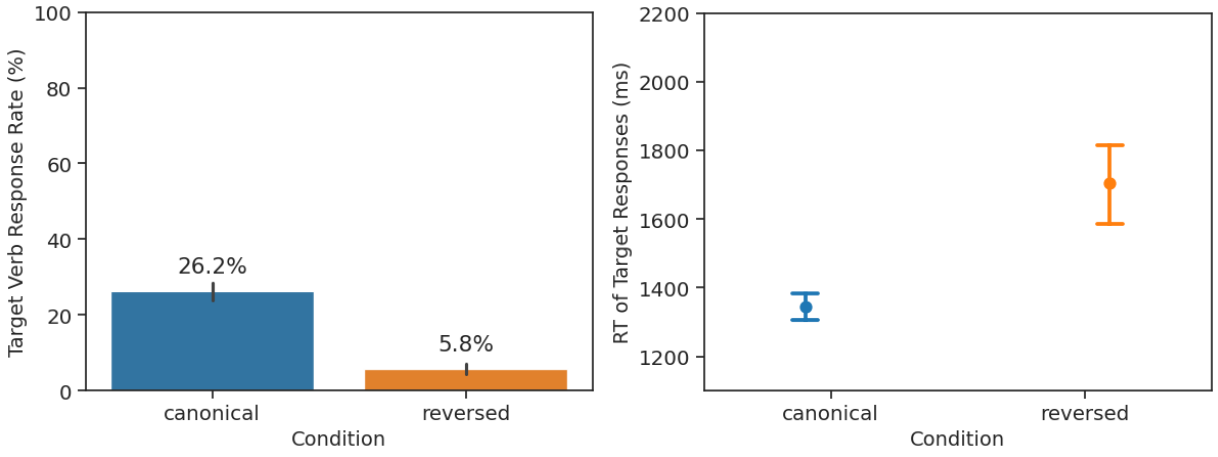


Figure 4: Percentages of target responses in the canonical and reversed conditions (left) and their mean RTs (right) in Experiment 2-1. Error bars represent 95% confidence intervals.

The cloze percentages and RTs of role-neutral responses indicate different candidate profiles between the canonical and reversed conditions (Figure 5). More role-neutral responses were produced in the reversed condition than in the canonical condition (β = 1.26, *SE* = .18, $p < .001$), while no differences were observed in RTs (Canonical: *M* = 1875 ms, *SD* = 523 ms; Reversed: *M* = 1805 ms, *SD* = 524 ms; $p > .05$). This indicates that more role-neutral responses were produced in the reversed condition than in the canonical condition despite having similar strengths. This pattern suggests a contrast in the strength of the field of candidates in the canonical and reversed contexts, such that there were weaker alternatives in the reversed condition than in the canonical condition.

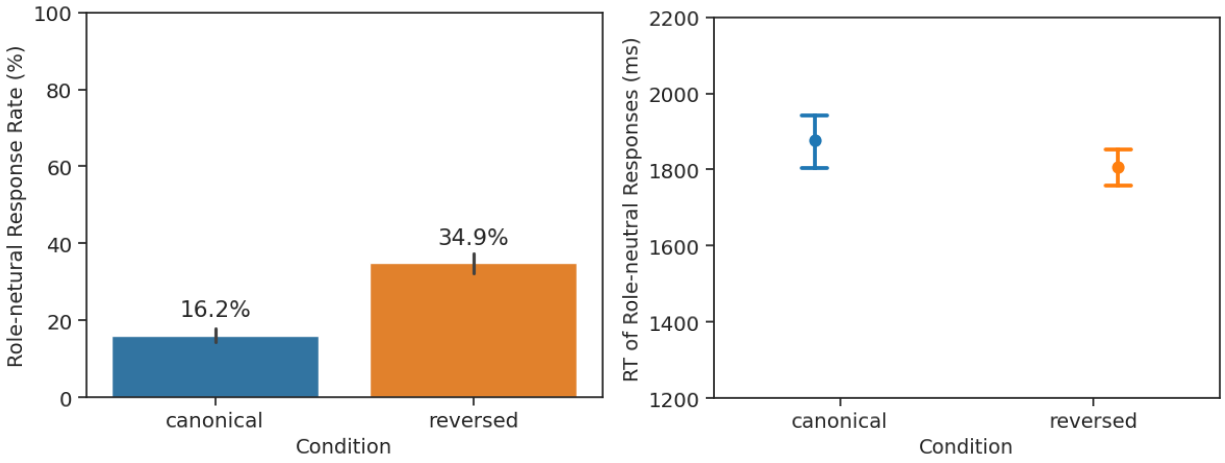


Figure 5: Percentages of role-neutral responses in the canonical and reversed conditions (left) and their mean RTs (right) in Experiment 2-1. Error bars represent 95% confidence intervals.

2.2.3 Discussion

The main goal of Experiment 2-1 was to test whether seeing role-reversed sentences during the experiment affected people's ability to produce role-appropriate responses in a speeded cloze paradigm. The question was whether being exposed to many role-reversed sentences in a comprehension experiment would make argument roles a less reliable cue to use for prediction and yield reduced role-sensitivity. If this were the case, we would expect to see an increase in role-

reversal errors in the hybrid comprehension-production experiment, where speeded cloze production trials were interleaved with comprehension trials, through which role-reversed sentences were presented to participants during the experiment.

The cloze response patterns showed that the exposure to many anomalous role-reversal sentences through interleaved comprehension trials did not affect role-sensitivity in the production trials. Cloze responses were overwhelmingly role-appropriate, and anomalous, role-reversed responses were rare, with a similar production rate as in a prior study using the same experimental materials without the interleaved comprehension trials (Nakamura, 2023). Moreover, when participants produced the same target verbs that were presented in the comprehension trials as cloze responses in the production trials, they produced them more often and with faster onset latencies in the (role-appropriate) canonical condition than in the (role-inappropriate) reversed condition. In order to get a better understanding of how these cloze response patterns translate into activation-based race dynamics, we examined the profiles of role-neutral responses, or verbs that were equally predictable by the preceding context. The analysis results indicated that role-neutral verbs had similar RTs in the canonical and reversed conditions but were more frequently produced in the reversed condition than in the canonical condition. This suggests that there was weaker competition among alternatives in the reversed context than in the canonical condition, since given the same speed, candidates would have a greater chance of winning the race when their competitors are weak (Staub et al., 2015). Despite the weaker competition, target verbs were produced less often, with slower speeds, in the reversed condition than in the canonical condition, indicating that they were not as strong candidates given the role-reversed context. These cloze patterns overall suggest that role information quickly and accurately constrained expectations for upcoming verbs

in the production trials, even when participants were exposed to many implausible, role-reversed sentences through the interleaved comprehension trials in the experiment.

While the results of Experiment 2-1 suggest that the comprehension-production divergence is unlikely to be driven by a difference in experimental materials, it is possible that participants in comprehension experiments and production experiments process sentence contexts differently because they engage in different tasks, not because they see different experimental materials. In Experiment 2-2, we examined whether putting participants in the same task setting where they must process the given contexts in the same way would elicit the same role-sensitivity in comprehension as in production.

2.3 Experiment 2-2

The same interleaved trials design used in Experiment 2-1 was used to put participants in a setting where they did not know at the start of each trial whether it was a comprehension or production trial, which ensured that participants were processing the sentences in a consistent manner while measuring N400 responses in the comprehension trials and speeded cloze responses in the production trials. We primarily focus on the N400 responses in the comprehension trials in Experiment 2-2, as the Experiment 2-1 speeded cloze results already indicated role-sensitivity in production in the interleaved trials experiment. If a significant N400 effect was observed in the comprehension trials (i.e., different N400 amplitudes between verbs in role-appropriate and role-reversed contexts), this would suggest that the change in the task is responsible for the change in role-sensitivity. If no N400 effect was found, like in previous EEG studies, it would indicate that even when people are engaged in the same task and processing the sentence contexts in the same way, they only show role-sensitivity in speeded cloze responses (Experiment 2-1) but not in N400

responses. This outcome would indicate that the comprehension and production measures reflect different parts of shared underlying processes.

2.3.1 Method

2.3.1.1 Participants

Twenty-three native English speaking adults (18-36 years old, mean age = 22) from the University of Maryland participated in the experiment. All were right-handed, had normal or corrected-to-normal vision and no history of neurological disorder. None of the participants had taken part in Experiment 2-1.

2.3.1.2 Materials

The stimuli were identical to Experiment 2-1.

2.3.1.3 Procedure

The same hybrid comprehension-production interleaved trials design from Experiment 2-1 was used in Experiment 2-2. The experiment was conducted in an EEG lab. Participants were seated in front of a computer and keyboard. A noise-canceling microphone was placed in front of the participant for recording speeded cloze responses in the production trials. The stimuli were presented using a Matlab script. The rest of the experiment procedure was identical to Experiment 2-1.

2.3.1.4 EEG recording and analysis

EEG was recorded from 30 active electrodes placed on an elastic cap (Electrocap International) according to the International 10-20 system. The vertical electrooculogram (EOG) was measured with two electrodes each placed above and below the left eye, and the horizontal EOG was recorded through electrodes placed at the outer canthus of each eye. EEG and EOG signals were

referenced to the left mastoid online and re-referenced to the average of the left and right mastoids offline. The AFz electrode was used as the ground electrode. The signal was digitized at a sampling rate of 500 Hz. Impedance was kept below 5 kΩ. The recordings were amplified and digitized online at 1000 Hz with a bandpass filter of 0.1-100 Hz.

The EEG data were analyzed in Matlab using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). The EEG data were time-locked to the target verb for the critical trials and the target noun for the control trials. A band-pass filter was applied at 0.1 Hz and 30 Hz, 12 dB/oct. The continuous EEG signal was epoched from 100 ms pre-stimulus to 800 ms after the stimulus onset, and baseline correction was applied using the pre-stimulus interval, from -100 ms to 0 ms. Artifact rejection was conducted by automatically detecting and rejecting epochs using a simple voltage threshold of -100 μV to 100 μV, and using a moving window peak-to-peak amplitude of 70 μV, where a 200 ms window was moved across the epoched data in 100 ms increments and any epoch with amplitude exceeding the set value was rejected. The resulting data were manually inspected to ensure that artifacts were successfully removed, and three participants with a rejection rate exceeding 40% were excluded from further analysis.

Statistical analysis was conducted following the recommendations from Luck (2022) and prior studies examining role-sensitivity in the N400 response (Liao et al., 2022). Mean amplitudes in the 300-500 ms time window from nine electrodes in the central-parietal area (C3, CZ, C4, CP3, CPZ, CP4, P3, PZ, P4) were subjected to a paired t-test, to observe the effect of context in the critical items (canonical vs. reversed), and the same analysis was carried out for the control items, to observe the effect of cloze (high-cloze vs. low-cloze).

2.3.1.5 Speeded cloze data analysis

Response rates and RTs in the production trials were examined in the same ways as in Experiment 2-1.

2.3.2 Results

2.3.2.1 Speeded cloze response rates and RTs in the production trials

Similar to the production results in Experiment 2-1, a strong majority of responses produced in the production trials in Experiment 2-2 were contextually appropriate given the preceding role context, and role-reversed responses were rare (Figure 6).

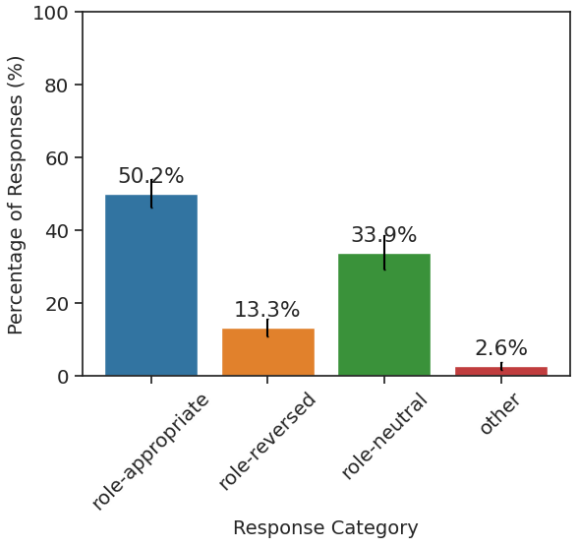


Figure 6: Percentages of cloze responses in each response category in Experiment 2-2. Error bars represent 95% confidence intervals.

For target response rates, there was a significant main effect of condition (β = -1.53, *SE* = .38, $p < .001$), where target responses were produced more in the canonical context than in the reversed context (Figure 7).

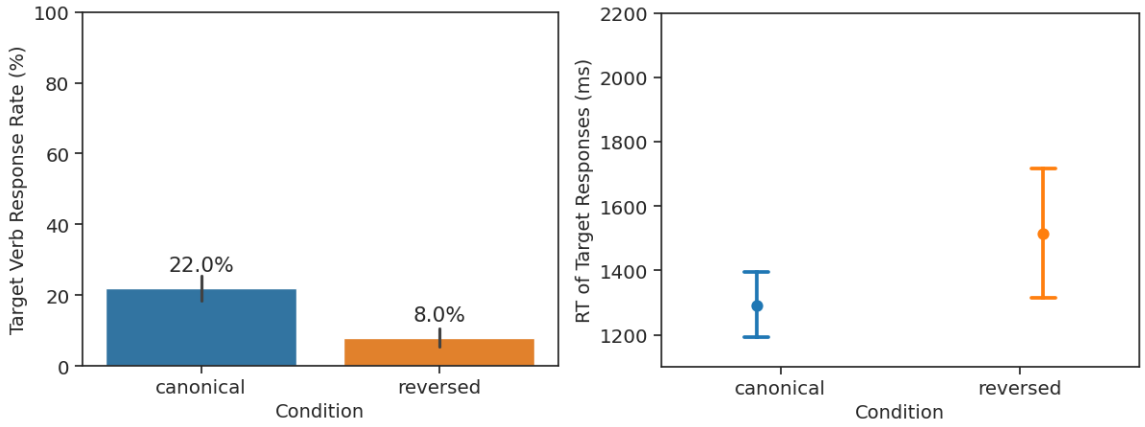


Figure 7: Percentages of target responses in the canonical and reversed conditions (left) and their mean RTs (right) in Experiment 2-2. Error bars represent 95% confidence intervals.

The response patterns of role-neutral verbs observed in Experiment 2-1 were also replicated in Experiment 2-2 (Figure 8). Role-neutral responses were produced more often in the reversed condition than in the canonical condition ($\beta = .98$, $SE = .27$, $p < .001$), with no differences in RTs between the two conditions (Canonical: $M = 1858$ ms, $SD = 714$ ms; Reversed: $M = 1908$ ms, $SD = 661$ ms; $p > .05$). This indicates that the strength of the field of candidates was weaker in the reversed condition than in the canonical condition. Target verbs nevertheless had lower cloze percentages and slower RTs in the reversed condition than in the canonical condition.

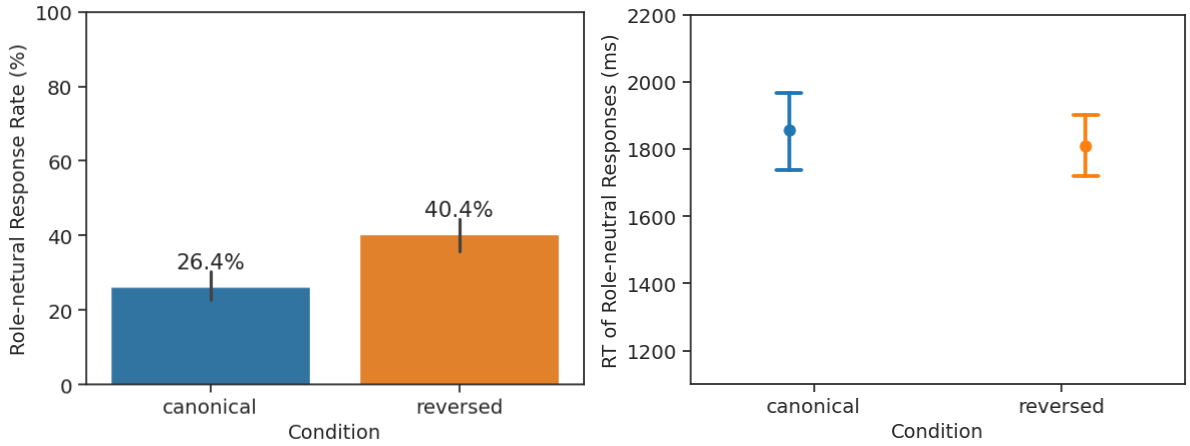


Figure 8: Percentages of role-neutral responses in the canonical and reversed conditions (left) and their mean RTs (right) in Experiment 2-2. Error bars represent 95% confidence intervals.

2.3.2.2 N400 responses in the comprehension trials

Participants judged the sentences presented in the comprehension trials as plausible with a higher rate in the canonical condition (93%) relative to the reversed condition (58%) [$t(19) = 7.55$, $p < .001$)], and in the high-cloze control condition (94%) relative to the low-cloze control condition (31%) [$t(19) = 10.67$, $p < .001$)].

The grand average N400 amplitudes in the canonical and reversed conditions for the critical items are shown in Figure 9. The comparison between the mean amplitudes from the canonical condition and reversed condition yielded a significant difference [$t(19) = 2.38$, $p = .028$], with a more negative N400 amplitude in the reversed condition ($M = 0.49$ µV, $SD = 3.15$ µV) than in the canonical condition ($M = 1.65$ µV, $SD = 2.84$ µV).

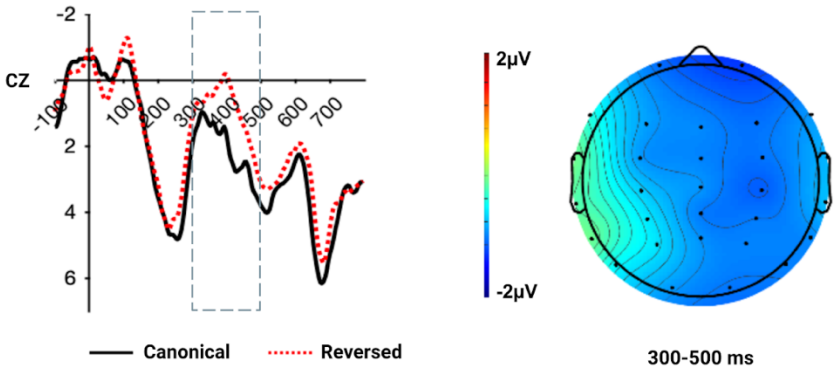


Figure 9: Grand average N400 amplitudes for the canonical and reversed conditions and the topographic distribution of the mean voltage difference at the Cz electrode in the 300-500 ms time window

The N400 amplitudes in the high-cloze and low-cloze conditions for the control items are shown in Figure 10. The comparison between the mean amplitudes from the high-cloze condition and low-cloze condition yielded a significant difference [$t(19) = 5.37$, $p < .001$], with a more negative N400 amplitude in the low-cloze condition ($M = -0.44$ µV, $SD = 1.77$ µV) than in the high-cloze condition ($M = 2.16$ µV, $SD = 2.18$ µV).

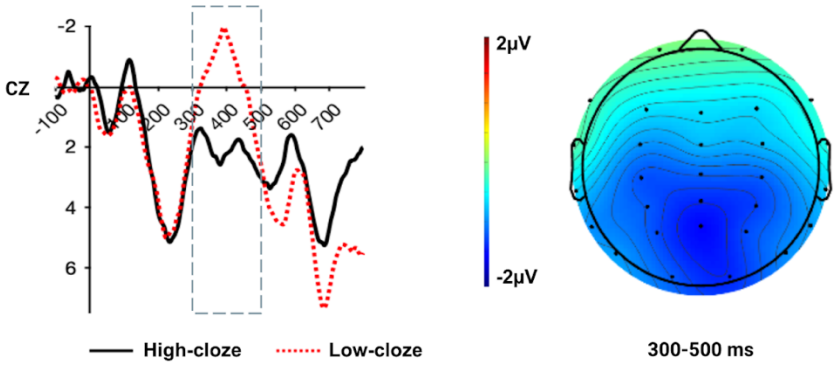


Figure 10: Grand average N400 amplitudes for the high-cloze (plausible) and low-cloze (implausible) control condition and the topographic distribution of the mean voltage difference at the Cz electrode in the 300-500 ms time window

2.3.3 Discussion

The goal of Experiment 2-2 was to test the hypothesis that the divergence between comprehension and production measures regarding argument role sensitivity arises from people being engaged in different tasks. The speeded cloze paradigm involves self-generating predictions and producing them aloud within a time limit, while a typical EEG experiment involves reading sentences presented word-by-word and occasionally making end-of-sentence judgments. The same interleaved trials design used in Experiment 2-1 was used in Experiment 2-2, where we measured EEG responses to target verbs presented in role-appropriate and role-reversed contexts (in the comprehension trials), while participants produced speeded cloze completions through interleaved production trials.

The results from the production trials support rapid use of argument roles in speeded cloze production. Responses were overwhelmingly role-appropriate, and role-reversed responses were rare, with a similar production rate as observed in previous speeded cloze studies (Chow et al., 2015; Nakamura et al., 2024) and in Experiment 2-1 of the current study. As in Experiment 2-1, the response patterns of target verbs and role-neutral verbs indicated that target verbs were weaker

candidates in the role-reversed context than in the role-appropriate, canonical context despite evidence suggesting that competition was weaker. These patterns repeatedly observed in speeded cloze responses in Experiments 2-1 and 2-2 suggest that argument roles have an immediate impact on production. They also reinforce the fact that seeing many role-reversal anomalies during the experiment does not reduce sensitivity to argument role information. People can rapidly use argument role information to produce role-appropriate verbs given a sentence context.

The key novel finding of Experiment 2-2 was that the EEG responses measured in the comprehension trials, which were interleaved with speeded cloze trials, revealed a significant N400 effect to role-reversal anomalies. The reduced N400 amplitudes for target verbs in the canonical condition relative to the reversed condition show that the target verbs elicited different neural responses in role-appropriate and role-reversed contexts. This indicates rapid sensitivity to argument roles during verb prediction, which stands in stark contrast to what has been found in previous EEG studies that repeatedly failed to find an N400 effect with role-reversals (e.g., Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Nakamura et al., 2024). While some studies have shown that manipulating the distance between the arguments and the verb (Chow et al., 2018) or the presentation rate of the context (Liao et al., 2022; Nakamura et al., 2024) can elicit an N400 effect during comprehension, the current experiment result shows that an N400 effect to role-reversals can appear without any manipulation of the context.

The results of Experiment 2-2 indicate that the comprehension-production contrast disappears when it is ensured that participants process the preceding contextual information in the same way. We found evidence for immediate use of argument role information to constrain expectations in both comprehension and production. In the General Discussion, we further discuss

the potential reasons for the emergence of immediate role-sensitivity in comprehension in the interleaved trials experiment.

2.4 General Discussion

The main goal of the present study was to examine whether people can quickly and accurately parse and use argument role information for later processes during real-time sentence comprehension and production. Previous studies have found that comprehenders initially show the same response to verbs presented in role-appropriate and role-inappropriate contexts, which has been taken as evidence for initial blindness to role information during comprehension (e.g., Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Stone et al., 2024; Kukona et al., 2011; Burnsky, 2022). On the other hand, production studies have found that people can quickly produce role-appropriate continuations and avoid role-reversal errors given sentence contexts (Chow et al., 2015; Nakamura et al., 2024). This comprehension-production contrast is unexpected given linking hypotheses for measures that are taken to reflect similar underlying activations of target words given a context, i.e., N400 and speeded cloze responses (Staub et al., 2015). To clarify the apparent contrast, we conducted two experiments in which comprehension and production trials were interleaved and presented in random order in the same experiment. This design ensured that participants did not know which kind of trial they were in at the start of each trial, making it possible to measure role-sensitivity in comprehension and production while participants processed the sentence contexts in the same way and saw the same materials in the same experiment setting. The results revealed immediate sensitivity to argument roles in both comprehension and production measures, indicating that argument role information has an immediate impact on

processing, when the task involves rapid use of contextual information to generate a sentence continuation.

The speeded cloze responses in the hybrid experiment showed evidence supporting rapid use of argument role information even when the experiment materials contained many role-reversal anomalies. Prior work has shown that predictive behavior in comprehension is affected by properties of the experiment materials, e.g., the proportion of predictable sentences presented during the experiment (e.g., Brothers et al., 2017; Lau et al., 2013; Ness & Meltzer-Asscher, 2021; but also see Van Wonderen & Nieuwland, 2023). Comprehension studies and production studies differ in terms of whether participants see role-reversed sentences during the experiment; comprehension studies involve presenting anomalous role-reversals, while (speeded cloze) production studies do not. The interleaved trials design used in the current study addressed this contrast by presenting role-reversals just as in a comprehension experiment and examining how it affected responses in the interleaved production trials. The results showed that the exposure to role-reversal anomalies did not change the production behavior. It did not hinder participants' ability to quickly and accurately provide sentence continuations that fit the preceding argument roles in the production trials. The vast majority of cloze responses were role-appropriate, and target verbs were produced more often, with faster onset latencies, in role-appropriate contexts than in role-reversed contexts, despite there being evidence for weaker competition among alternatives in the reversed contexts which could allow role-reversals to get produced more often (Staub et al., 2015). These results indicate that role information was actively used to constrain speeded cloze responses, even when a portion of the experiment trials contained role-inappropriate sentence continuations. Seeing many instances of role-reversals, therefore, did not make the production behavior resemble the role-insensitivity observed in comprehension, suggesting that the absence

of role-reversal materials in production experiments is unlikely the main cause of the apparent comprehension-production asymmetry.

While the interleaved trials design yielded a consistent finding in production, i.e., rapid use of argument roles, it resulted in a significant change in comprehension. Target verbs elicited smaller N400 amplitudes when presented in role-appropriate contexts compared to role-reversed contexts, indicating immediate sensitivity to argument roles. Two key aspects of this finding are noteworthy. One is that the same sensitivity was observed in production and comprehension measures when participants were in the same experiment setting and performed the same task. The interleaved trials design used in the current study ensured that the sentence contexts were processed in the same way in the comprehension and production trials, since participants did not know at the beginning of each trial whether it was a comprehension trial or a production trial. Using this task, we found evidence for rapid use of argument roles in both comprehension and production measures, indicating that the apparent contrast observed across prior studies was mostly driven by participants in comprehension experiments and production experiments engaging in different tasks and processing the given sentence contexts differently in those experiments.

Another key aspect to the comprehension results is that immediate role-sensitivity was observed in a comprehension measure in the hybrid interleaved trials design. The emergence of the N400 effect in response to role-reversal anomalies stands in contrast to the absence of the N400 effect, as repeatedly reported in previous comprehension-only paradigms (e.g., Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Stone et al., 2024). While some studies have found sensitivity to argument roles in the N400 measure with an additional manipulation of time (Chow et al., 2018) or experiment materials (Ehrenhofer et al., 2019), this is the first time where the N400

effect has been observed using the same materials that previously failed to elicit N400 role-sensitivity (Chow et al., 2016). The main difference between a comprehension-only experiment and our hybrid comprehension-production task is the presence of (speeded cloze) production trials. Unlike in comprehension-only experiments, participants in the hybrid task had to produce sentence continuations in a portion of trials.

The N400 effect found in our hybrid experiment appears to resemble prior findings of increased predictability effects in tasks that involved production. For example, Hintz et al. (2016) found that interleaving picture-naming production trials with self-paced reading trials yielded predictability effects in reading times in the self-paced reading trials which were not observed in the absence of the picture-naming trials. Lelonkiewicz et al. (2021) found that reading aloud sentence contexts led to faster sentence-final word recognition times relative to reading the contexts silently. The authors claimed that in the read-aloud condition, as opposed to the silent reading condition, participants engaged their production system which facilitated prediction and yielded faster processing of predictable sentence-final words. More generally, it has been claimed that comprehenders can use the production system to predict upcoming input during comprehension (Pickering & Garrod, 2013; Gambi & Pickering, 2013; Dell & Chang, 2014; Pickering & Gambi, 2018; Martin et al., 2018). According to this view, prediction by production occurs through a process of covertly imitating the given context through the comprehender's own production system and generating predictions for upcoming input based on the comprehender's understanding of the intended message (Pickering & Gambi, 2018).

Despite the resemblance, there are notable differences between the prior studies and the current study which motivate a more specific account than claiming that prediction was facilitated by the recruitment of the production system. Reference to production masks a key difference

between the tasks used in the prior studies and our hybrid paradigm, since production involves i) planning what to say and ii) engaging articulatory processes. The production tasks used in the earlier studies involved targeted articulation of given words or sentence contexts; participants did not have to choose what to say or generate their own responses. The hybrid task in our study involved planning what to say and generating one's own sentence continuation, and in the critical trials no articulation was required. Second, our findings reflect a qualitative change in sensitivity (i.e., responses to role-appropriate and role-inappropriate verbs), rather than mere facilitation of processes that were occurring even in the absence of the production task (e.g., faster reading times for predictable words). Therefore, given the specific properties of our hybrid paradigm, a suitable account must capture how the engagement of the planning mechanism, in particular, led to a qualitative change in sensitivity of role-based expectations.

What could cause comprehenders to show immediate role-sensitivity in our hybrid interleaved trials design, in contrast to comprehension-only experiments? According to the view that blames the speed at which role-based verb expectations are generated (Liao et al., 2022; Chow et al., 2018; Nakamura et al., 2024), immediate role-sensitivity could arise if the processes that cause the delayed sensitivity are facilitated in the hybrid task. Alternatively, based on the view that argument roles have immediate but erroneous effects on prediction because of stronger influences from prior knowledge of event probabilities, independent of the given argument roles (Kim et al., 2015; Kuperberg, 2016; Li & Ettinger, 2022; Rabovsky et al., 2018; Stone et al., 2024), explaining the current findings would require claiming that the hybrid task triggered a change in the way different sources of information are used to generate representations. This might involve a suppression of event knowledge or a reweighting of cues such that role information can successfully constrain predictions.

While further empirical investigation is needed to determine which of the above is responsible for the change in the N400 response in the current study, we are sympathetic to the view that the hybrid task facilitates already ongoing processes rather than invoking a fundamental change in the way representations are generated. Assuming a distinction between i) generating representations and ii) committing to a single representation (which could, in principle, be sent for articulation), the hybrid task would require more engagement of the second process. This is because the interleaved speeded cloze trials require rapid commitment to a representation, in order to produce a single sentence continuation. Then, rather than necessarily involving a separate production system, or an inherently different processing strategy, the hybrid task could engage processes in comprehension that are more necessary in production. That is, in a situation where there is no pressure to select a single best guess for the upcoming input, comprehenders may entertain multiple possible representations simultaneously (Roland et al., 2012; Staub et al., 2015; Frisson et al., 2017; Ness & Meltzer-Asscher, 2021) without quickly committing to a particular candidate. Conversely, in a situation where there is a benefit to selecting the most probable representation as quickly as possible, as in production where a single candidate must be sent for articulation, processes that can help to quickly reach a state with a single representation could be more engaged. Importantly, these processes are engaged not only in the production trials but also in the comprehension trials in the hybrid task, as participants do not know at the start of each trial whether they will be given a sentence continuation or they will have to produce their own continuation. Hence, they would recruit the same processes that are needed to produce a sentence continuation under time pressure even when the trial eventually turns out to be a comprehension trial, in which role-sensitivity is measured through the N400 response to presented target words.

We propose two mechanistic ways in which the need to quickly commit to a single representation could yield immediate role-sensitivity. They are both based on modulations of an account of the timing effect reported in prior studies: role-sensitivity in comprehension emerges with additional time (Chow et al., 2018; Stone et al., 2024; Nakamura et al., 2024). Under the view that time is one key factor that determines role-sensitivity, accelerating parts of the underlying processes could capture the role-sensitivity observed in the current study.

One possibility is that rapid commitment facilitates the process of filtering out role-inappropriate candidates that were erroneously generated by the context. This assumes that initial verb predictions are role-insensitive, and a subsequent monitoring process filters out role-inappropriate candidates (Nakamura et al., 2024). Accelerating the process of inhibiting role-inappropriate verb candidates would yield the N400 effect observed in the hybrid task, i.e., immediate sensitivity to argument roles.

Alternatively, it is possible that the generation of verb candidates based on context, i.e., what occurs before the verb is presented, is role-sensitive, but that processes that occur after the verb is presented are more vulnerable to role-reversals. Recently, Lee & Phillips (in prep.) tested the specific contribution of confrontation of role-reversal verbs by presenting distractor verbs while participants were engaged in a speeded cloze task. They found that participants were particularly susceptible to errors when presented with role-reversal distractors. This recreated in production a role-insensitive profile similar to that typically seen in comprehension. This highlights the critical role of confrontation with inappropriate verbs to role-insensitivity. Second, in analyzing variability across individual items in the study, the authors found that vulnerability to role-reversal distractors was much lower in items where there was a readily accessible role-appropriate continuation. This highlights that the availability of good role-appropriate

continuations has a protective effect in making participants invulnerable to reversals. Hence, a possible impact of the hybrid paradigm in the current study is that it may have accelerated the generation of role-appropriate candidates, leading to the emergence of role-sensitivity in the N400 response in the comprehension trials.

2.5 Conclusion

The present study aimed to examine an apparent comprehension-production contrast with regard to whether people show immediate sensitivity to given argument role information during real-time sentence processing. Previous comprehension studies have shown that initial verb expectations are not constrained by argument roles, while production studies have shown that they are rapidly used in generating next-word continuations. By using a hybrid interleaved trials design, we measured role-sensitivity during comprehension and production while participants were put in an identical task setting and processed the given context in the same way. This yielded immediate role-sensitivity in both comprehension (EEG) and production (speeded cloze) measures, indicating that the contrast found in prior work was driven by different task demands in different experiment settings. Engaging in a task that requires active generation of candidates and commitment to a single representation can yield immediate sensitivity to argument roles during real-time processing.

Chapter 3: Confronting words in the absence of strong alternatives

3.1 Introduction

A large body of psycholinguistic work has shown that comprehenders actively generate multiple possible representations based on information made available in preceding sentence contexts. Studies have revealed that people build complex structures and create expectations for what kinds of inputs are a good fit with the representation that they are currently building, given previous information. For example, comprehenders expect a certain category of verbs given two preceding arguments, as demonstrated in filler-gap processing (e.g., Omaki et al., 2015).

In contrast to the large body of work indicating rapid and accurate generation of possible representations during sentence comprehension, the role-reversal phenomenon has presented a challenge against the idea of successful generation processes. Many studies have shown that comprehenders fail to immediately distinguish between verbs that appear in role-appropriate and role-reversed contexts (a vs. b).

a) *…which customer the waitress had* ***served***

b) *…which waitress the customer had* ***served***

This phenomenon has led to accounts of predictive processing that claim that generation of expectations is not always immediately context-sensitive and that sometimes comprehenders generate contextually inappropriate representations, i.e., 'faulty generation.' We will use the term, 'faulty generation' as a cover term that includes the generation of any type or level of representation that is not fully consistent with the given preceding information. This includes generating role-neutral, lexically associated candidates due to limited time to generate role-specific candidates (Chow et al., 2016; Liao et al., 2022) or to filter out role-inappropriate candidates (Nakamura et al., 2024). It also includes generating role-inappropriate candidates due to semantic

cues outweighing syntactic cues (Kuperberg et al., 2016; Rabovsky et al., 2018; Stone et al., 2024), processing the preceding context using heuristics based on distributional information about more or less probable events (Li & Ettinger, 2023).

Previous studies have largely blamed processes that occur before a verb is presented in the input for the initial blindness to role-reversal anomalies. However, this notion of faulty generation is directly challenged by recent speeded cloze production studies that showed that when asked to rapidly generate sentence continuations, speakers' responses are overwhelmingly role-appropriate (Chow et al., 2015; Nakamura et al., 2024). That is, speakers demonstrate the capacity to rapidly generate and produce sentence continuations that are fully consistent with the argument roles assigned in the preceding context.

In order to resolve the conflicting findings between comprehension and production measures, Lee & Phillips (under review) tested whether the contrast might be explained by the different task demands associated with speeded cloze and comprehension tasks. The authors showed that rapid commitment to a single response could facilitate role-sensitive generation processes. The study showed that it is possible to elicit in comprehension the same role-sensitivity found in production by putting people in a similar setting that requires rapid commitment to a single representation. This suggests that the difference observed between comprehension and production may not necessarily arise from engaging in fundamentally different mechanisms but rather by engaging in the same underlying processes to different extents depending on task-related goals. While this work shows that shifting the vulnerability in comprehension through a different task is possible, it remains a question why in previous comprehension studies, the same role-sensitivity is not observed as when people are encouraged to quickly commit to a single response, in a hybrid paradigm used in Lee & Phillips (under review).

In the current study, we focused on the question regarding to what extent the vulnerability comes from processes that occur before the verb appears in the input, as opposed to processes that are triggered once a verb is encountered. Speeded cloze response patterns demonstrate that people generate role-appropriate verb predictions, which contrasts with the role-insensitivity reflected in the N400 response during comprehension. Comprehension experiment paradigms (EEG/ERP experiments, the visual world paradigm, self-paced reading, reading while eye-tracking, etc.) differ from the speeded cloze paradigm in that in comprehension experiments, participants explicitly see a target word, to which responses are then measured. In EEG experiments, for example, participants are presented with sentences including target words, and inferences are made about the underlying mechanism that led to the N400 response to those target words, based on participants' responses to them. With role-reversal anomalies, participants are presented with role-reversal verbs, like *served* in b), and the N400 response is measured and compared with the N400 response to the same verb presented in the role-appropriate context, a). The critical challenge is determining the status of a target word before and after participants confront it in the input. While the N400 response to a target input is often taken to reflect the status of the target before it was presented, it is also possible that additional processes occur once the target word is encountered in the input. In contrast, speeded cloze production does not involve any explicit presentation of role-reversal verbs, as it is the participants that provide their own sentence continuations. It is possible that confronting a role-reversal verb evokes a particular response to the verb during comprehension which does not directly reflect the original status of the verb when it was not explicitly confronted in the input.

The present study aimed to pinpoint the vulnerability that gives rise to role-inappropriate interpretations, by teasing apart processes that occur before and after a word is encountered in a

sentence context. Separating the two processes allows us to pinpoint which stage of the process is more vulnerable to role-reversal anomalies. This in turn, can tell us whether the internal generation processes are susceptible to contextually inappropriate representations, or the processes that occur once a word is encountered are susceptible to role-inappropriate interpretations.

We tested whether presenting role-appropriate and role-inappropriate verbs during a speeded cloze task such that it resembles the way responses are elicited and measured in a comprehension task would affect the role-sensitivity typically observed in production. We designed a speeded cloze interference paradigm, which involved presenting distractor verbs in role-appropriate and role-inappropriate contexts, immediately before participants produced a response in a speeded cloze task. Speakers were instructed to either produce their own continuation or the distractor if the distractor was a good fit to the preceding context.

The interference paradigm operates on a similar logic used in picture-word interference tasks that have been widely used to study lexical access in production. Studies have shown that a semantically related distractor superimposed on a target picture interferes with the naming of the picture (e.g., Schriefers et al., 1990). In our speeded cloze interference paradigm, a distractor verb could, in principle, interfere with the production of a cloze response and result in the production of the distractor itself if it was considered a better sentence continuation than one that was self-generated based on the prior context. If seeing a role-reversal distractor as opposed to an unrelated distractor elicits more role-reversal errors in this distractor paradigm, it would suggest that at least part of the role-insensitivity in comprehension arises once the verb is encountered in the input, rather than before. If we see any effect of the distractor on the response rates and latencies of role-reversal verbs, it would suggest that confronting those verbs in the input causes an originally role-sensitive generation process to yield reduced role-sensitivity. This kind of result would challenge

the claim that the lack of role-sensitivity found in previous EEG studies is mainly driven by the comprehender actively generating role-inappropriate candidates based on preceding contexts and put greater emphasis on processes that occur once the verb is encountered.

3.2 Experiment 3-1

The goal of Experiment 3-1 was to test whether the role-sensitivity typically observed in speeded cloze production would remain robust even when speakers confront a role-reversal verb as a sentence continuation, similar to a comprehension setting. We examined i) how much confronting a target distractor, relative to an unrelated control distractor, led to an increase in the target verb's cloze probability, and ii) whether confronting a target distractor affected the cloze probability and RTs of the target verb in role-appropriate and role-reversed contexts equally. The comparison between the two contexts would indicate role-(in)sensitivity, e.g., a higher rate of target distractor responses in the role-appropriate context than in the role-reversed context would indicate role-insensitivity. The comparison between the two distractor types would reveal any effect of confrontation, e.g., a higher rate of target distractor responses when preceded by target distractors than control distractors would indicate an increase in production due to confrontation. Finally, the interaction between context and distractor type would indicate whether confrontation affected target responses differently in the role-appropriate and role-reversed contexts, e.g., a larger increase in target distractor responses following target distractors than control distractors in the role-reversed context than in the role-appropriate context would indicate that confrontation led to a greater increase in role-reversal errors than role-appropriate responses.

3.2.1 Method

3.2.1.1 Participants

Fifty-five native speakers of English on Amazon Mechanical Turk participated in the experiment.

3.2.1.2 Materials

Experiment stimuli were adapted from Chow et al.'s (2016) EEG study. The critical items in that study were 60 pairs of sentences, each with a canonical or reversed argument role context, followed by a target verb that was role-appropriate in the canonical context and role-inappropriate in the reversed context. For our cloze interference paradigm, we truncated these sentences prior to the target verb and used them as contexts to elicit cloze continuations. The last word in the context for the critical items was always "had," which was presented in red font, indicating that a continuation should be produced following the word. The sentence contexts, both canonical and reversed, each appeared in two different distractor conditions. Each sentence context appeared either with a target distractor or a control distractor. The target distractors were the same target verbs used in Chow et al.'s (2016) EEG study, which did not find immediate N400 role-sensitivity. The control distractors served as a baseline condition, in order to examine participants' behavior when an unrelated and implausible distractor was presented following each sentence context. The control distractors were selected from a list of 1,000 most frequent verbs in the Corpus of Contemporary American English (Davies, 2008-). We randomly sampled a verb and inserted it as a completion for each sentence fragment and checked to ensure that they were implausible completions given the preceding arguments in both contexts. An example set of critical items is presented in Table 2. In addition to the critical items, the same filler items were taken from Chow et al. (2016). The same target verbs in the prior study were used for the target distractors, and

nonwords were used as the control distractors for the filler items. This was to prevent participants from using a strategy of simply relying on the presented distractor to produce a cloze continuation.

Table 2: Experimental conditions and sample set of critical items in Experiments 2-1 & 2-2

Context		**Distractor Type**		
The restaurant owner forgot…		Target	Control	Weak
Canonical	*…which customer the waitress had*	served	thrown	hated
Reversed	*…which waitress the customer had*			

Note. Weak distractors were only included in Experiment 3-2. Target distractors were 'role-reversal verbs' when presented in the reversed context.

In total, each participant saw 240 sentence fragments (120 critical items and 120 filler items), where 60 critical items appeared in the canonical context and the other 60 critical items appeared with the reversed context, and for each context, 30 items were presented with target distractors and the other 30 items were presented with control distractors.

3.2.1.3 Procedure

The speeded cloze interference paradigm was administered online using PCIbex. Each trial began with a fixation mark '+' (800 ms), then the sentence fragment presented in RSVP format with a 530 ms SOA (300 ms per word, 230 ms blank between words), followed by the last word of the sentence fragment, which appeared in red font (300 ms), indicating that the participant should produce a continuation. This was followed by the distractor word (300 ms) presented between a front mask (150 ms) and a back mask (3000 ms), all in blue font. A series of #'s equivalent to the number of letters of the distractor word was used for the masks. Recording began when the front mask appeared. The back mask stayed on the screen while participants produced a cloze response,

within a 3-second time limit. Participants pressed the spacebar to proceed to the next trial. An illustration of a trial is displayed in Figure 11.

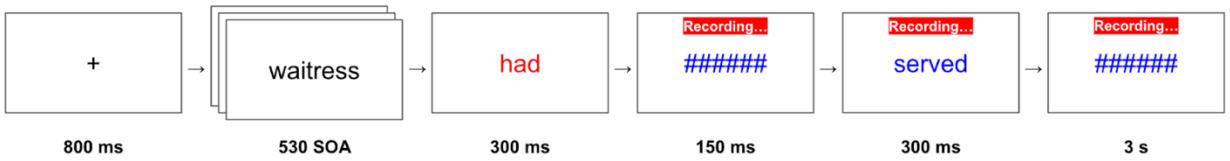


Figure 11: Example presentation of a trial in the speeded cloze interference paradigm. The sentence context is presented word-by-word in the center of the screen. The final word of the sentence fragment is presented in red font, which serves as the cue to produce a cloze response. The distractor word is presented immediately after the final word, with a front and back mask, all in blue font. Participants had three seconds to produce a response. Recording began following the final word, along with the presentation of the front mask.

Participants were asked to provide a likely continuation of each sentence fragment, after they saw a word in red font. They were instructed that they would see a blue distractor word and to produce the word if it was the word they were planning to produce and to ignore it otherwise. Participants were encouraged to respond as quickly and accurately as possible and were informed about the time limit. A step-by-step introduction of the paradigm including examples of good and bad responses, with four practice trials each with and without the distractor presentation was included in the beginning of the experiment. The entire experiment took about an hour to complete. All participants gave written consent before taking part in the experiment and received monetary compensation at the end of the experiment.

3.2.1.4 Analysis

The recorded speech data were pre-processed using a pipeline developed and made publicly available by the first author: https://github.com/ekrosalee/SpeechDataProcessingPipeline. The pipeline uses the speech-to-text API provided by AssemblyAI (https://www.assemblyai.com) to obtain transcriptions of the recorded speech data and uses Chronset, an automated speech onset detection tool (Roux et al., 2017), to obtain speech onset latencies. The transcriptions and timing measures are then manually inspected and adjusted using Praat (Boersma, 2001).

Trials with incomplete or incomprehensible recordings were excluded from analysis. With the remaining 6154 responses, we first examined the trials where participants did not produce the distractor verb and instead produced an alternative response. We examined how confronting a target or control distractor affected the proportions of alternative non-distractor responses. These 'non-distractor' responses were coded using the same method as in prior studies (Nakamura et al., 2024; Lee & Phillips, under review). Each response that was produced in any condition was marked as being a better fit in the canonical role order, the reversed role order, or (n)either. The responses were then marked as 'role-appropriate' if they were produced in the context with the preferred argument roles, 'role-reversed' if they were produced in the context with the opposite roles, 'role-neutral' if there was no clear preference for a particular role order, and 'other' if they were not verbs.

Second, in order to examine how confronting a distractor affected the verb's cloze probability in role-appropriate and role-reversed contexts, we calculated cloze percentages and RTs of target verbs in each condition and compared across conditions, between contexts (canonical vs. reversed) and between distractor types (target vs. control). The same analyses were conducted with trials where participants produced a control verb as the cloze response. The control verb analysis served the purpose of checking whether participants successfully performed the task and avoided producing the distractor when it was an unrelated and implausible continuation. We expected to see a low rate of control verb responses across conditions, including when the distractor was a control verb.

A higher rate of target responses in the canonical context than in the reversed context would indicate role-sensitivity. Critically, target response rates were examined in relation to distractor type. Comparing target verb response rates between trials where participants saw the target verb

as the distractor or an unrelated control distractor would indicate how likely confronting a verb increases its production rate, i.e., the effect of confrontation. For example, seeing a large increase in role-reversal errors after seeing role-reversal distractors, relative to control distractors, would indicate reduced role-sensitivity caused by processes that occur after seeing a role-reversal verb.

Statistical analyses of target and control verb response rates were each conducted with a Bayesian generalized linear mixed-effects model using the *brms* package (v2.21.0; Bürkner, 2017) in R (v4.3.2; R Core Team, 2021). Models with a binomial distribution and logit link were constructed for response rates, and models with a lognormal likelihood were constructed for RTs. The models included context, distractor type, and their interaction as fixed effects and random intercepts and slopes for all effects by participants and items. Models for the RTs of control distractor responses exceptionally included context as the only fixed effect, without distractor type, as there were no control verbs produced following a target distractor. The fixed effects were entered in the models with sum contrast coding (canonical: −0.5, reversed: 0.5; target: −0.5, control: 0.5). For all models, we used weakly informative priors (Gelman et al., 2008), where priors for the intercepts and regression estimates were set to N(0, 1) and default priors for all others. We ran four sampling chains of 4,000 iterations each, using the first half of the iterations for warmup. The parameter estimates and 95% credible intervals (CrIs) are reported on the log-odds scale for response rates and on the log-millisecond scale for RTs. We indicate that the parameter has a reliable effect if zero is not within the 95% CrI, which means that there is a 95% probability that the parameter lies within the interval.

3.2.2 Results

3.2.2.1 Distribution of non-distractor responses

The distribution of non-distractor responses is presented in Figure 12

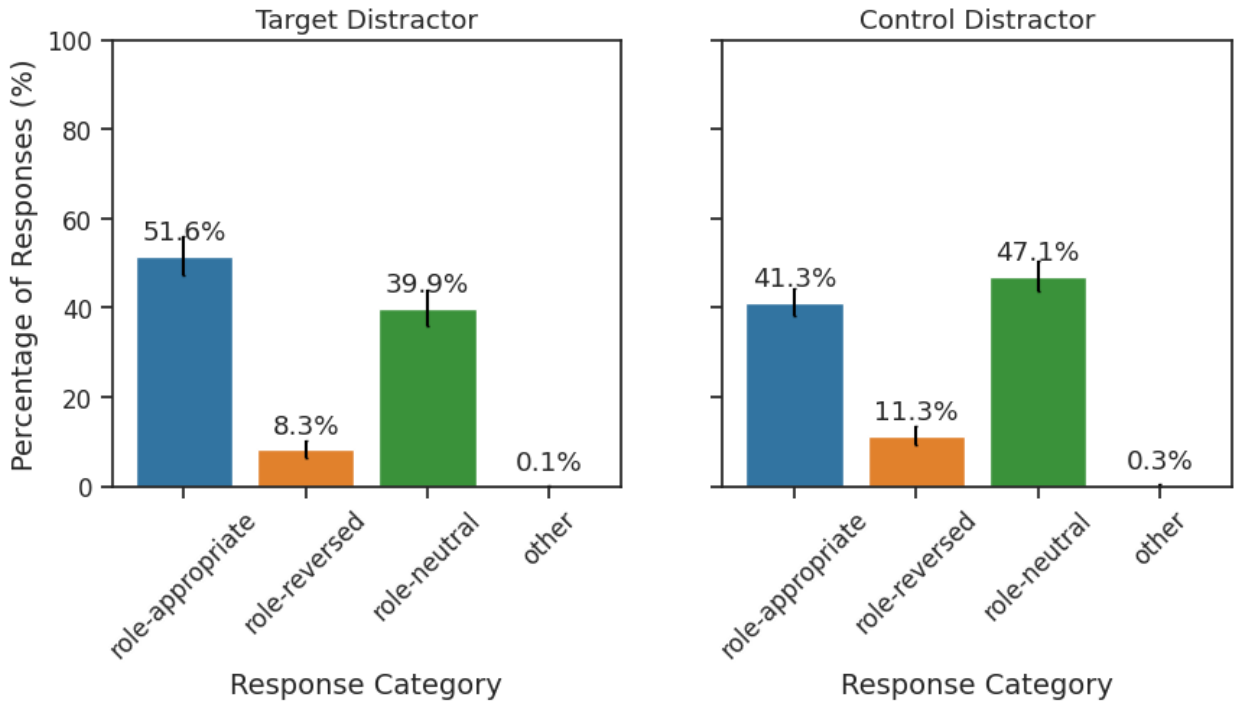


Figure 12: Distribution of responses excluding distractor verbs.

The vast majority of non-distractor responses were role-appropriate, with a low rate of role-reversal errors, similar to what has been found in prior speeded cloze studies without the distractor paradigm (Nakamura et al., 2024; Lee & Phillips, under review). This suggests that the confrontation of distractors has a minimal impact on the distribution of responses, regardless of whether the distractor is a tempting lure (target) or an unrelated word (control).

3.2.2.2 Cloze percentages and RTs of distractor responses

The cloze percentages of target and control verb responses are plotted in Figure 13.

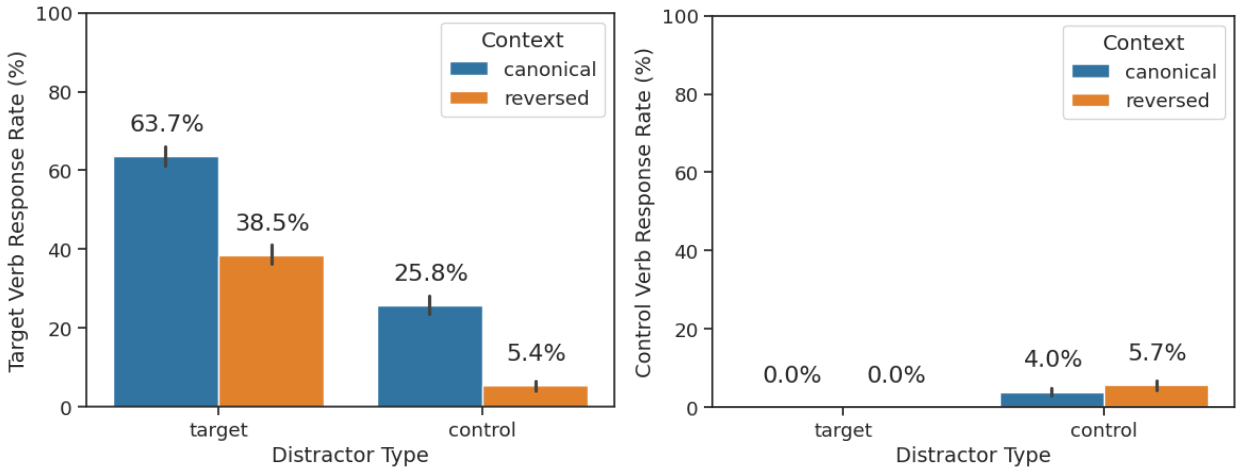


Figure 13: Experiment 3-1 cloze percentages of target verbs (e.g., *served*) (left) and control verbs (e.g., *thrown*) (right) in each condition. Error bars indicate 95% confidence intervals.

Figure 13 (left side) represents the percentage of target verbs produced in each condition. With control distractors, target verbs were rarely produced in the reversed context, less often than in the canonical context, with rates comparable to what has been observed in prior speeded cloze studies without any distractors. This indicates that confronting unrelated verbs did not significantly affect context-based generation processes which have been shown to be role-sensitive. Crucially, target distractors elicited a dramatic increase in target verb responses in both canonical and reversed contexts. Target verb cloze percentages increased from 26% to 64% in the canonical context, and notably, from 5% to 39% in the reversed context, indicating a large increase in role-reversal errors. The model for target verb response rates showed strong evidence for an effect of distractor type (β = −2.24, 95% CrI [−2.59, −1.89]), where more target verbs were produced following target distractors than control distractors. The model also revealed strong support for an effect of context (β = −1.68, 95% CrI [−2.00, −1.37]), where more target verbs were produced in the canonical context than in the reversed context. There was also evidence for an interaction between context and distractor type (β = −0.67, 95% CrI [−1.10, −0.25]). Nested models splitting the data into canonical and reversed contexts indicated a greater posterior mean of the effect of

distractor type in the reversed context (β = −2.44, 95% CrI [−2.23, −1.54]) than in the canonical context (β = −1.88, 95% CrI [−2.85, −2.01]).

Control verbs were not produced following target distractors and rarely produced following control distractors (Figure 13; right side), indicating that participants successfully avoided producing the unrelated control verbs as a cloze response. The model for control verb response rates indicated an effect of distractor type (β = 3.01, 95% CrI [2.08, 4.00]), indicating that control verbs were more frequently produced following control distractors than target distractors. There was no suggestive evidence for an effect of context (β = 0.30, 95% CrI [−0.49, 1.09]) nor an interaction between distractor type and context (β = 0.29, 95% CrI [−1.07, 1.66]).

The mean RTs of target and control verb responses are plotted in Figure 14.

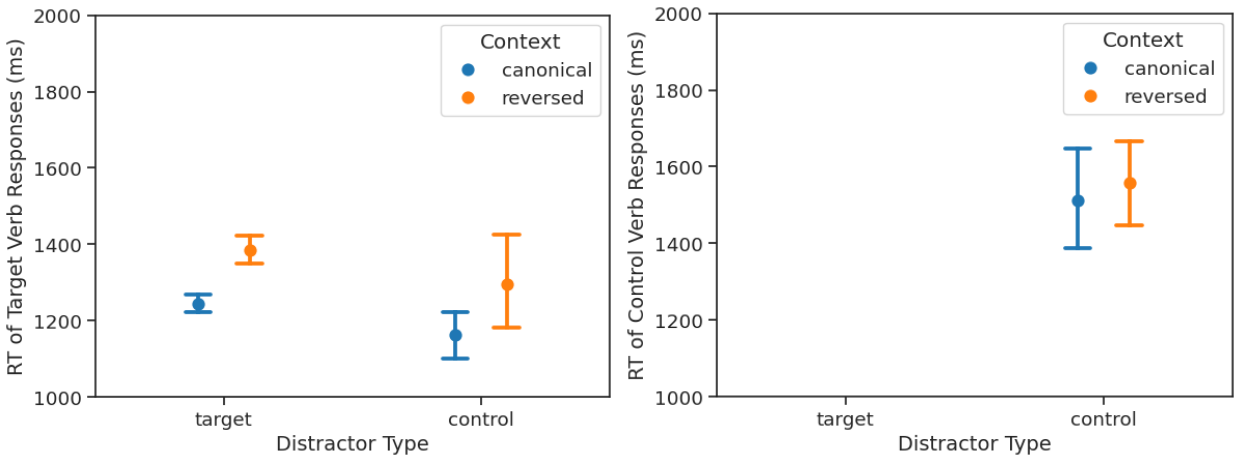


Figure 14: Experiment 3-1 cloze RTs of target verb responses (e.g., *served*) (left) and control verb responses (e.g., *thrown*) (right) following target and control distractor presentation. Error bars indicate 95% confidence intervals.

RTs of target verb responses showed significant differences across conditions (Figure 14). The model was consistent with an effect of context (β = 0.15, 95% CrI [0.09, 0.20]), where target verbs were produced with slower RTs in the reversed context than in the canonical context. Although target verb responses appeared to be overall slower following target distractors than control distractors, the model did not indicate strong evidence for an effect of distractor type (β =

−0.04, 95% CrI [−0.11, 0.03]) or an interaction between context and distractor type (β = 0.05, 95% CrI [−0.05, 0.14]).

RTs of control distractor responses were comparable between the canonical and reversed contexts, and the model indicated there was no evidence for an effect of context (β = 0.00, 95% CrI [−0.09, 0.10]).

3.2.3 Discussion

The main goal of Experiment 3-1 was to test whether confronting role-reversal verbs in the input would elicit a reduced role-sensitivity in speeded cloze production, a measure that has been shown to yield robust role-sensitivity (Chow et al., 2015; Nakamura et al., 2024). We designed a speeded cloze interference paradigm, where in each trial, participants saw a predictable role-appropriate verb, low-cloze role-reversal verb, or an unrelated verb as a distractor as a continuation of a sentence context, immediately before they themselves had to produce a cloze response. The rate and speed at which participants produced the presented distractor were measured and compared across conditions.

The overall distribution of responses revealed high role-sensitivity in the alternative responses participants produced, in lieu of the presented distractor. The majority of responses were either specific to the presented argument role order or role-neutral, meaning it fit equally well in either order. Role-reversal responses were overall rare. This indicates that even with the distractor confrontation, the responses that participants eventually produced when they successfully avoided the distractor were largely role-sensitive. This reinforces the role-sensitivity in quickly generating next-word continuations given the preceding context.

Response rates and RTs of distractor verbs revealed that participants were more likely to produce a target verb when presented as a distractor, relative to when confronted with an unrelated

distractor. This increase in target response rates was greater in the role-reversed context than the canonical context, indicating that participants were highly tempted to produce a presented role-reversal verb when it was explicitly presented as a continuation. The target distractor presentation did not affect the speed of responses, as responses were equally slower for role-reversal responses than role-appropriate responses, regardless of which distractor participants saw. This suggests that

A remaining question is to what extent the processes that occur after distractor presentation are role-sensitive. Given the large increase in role-reversal errors after confrontation, it is possible that the evaluation of a given input is highly role-insensitive. That is, participants might erroneously judge the target distractor in the reversed context as an appropriate continuation, because they do not actively use the preceding argument role information during the evaluation process. In order to examine the role-sensitivity of post-confrontation processing, we carried out another experiment which tested whether a similarly low-cloze response but role-appropriate would elicit a similar increase in cloze probability following distractor confrontation.

3.3 Experiment 3-2

In Experiment 3-2, we examined role-sensitivity after confrontation of a role-appropriate and role-reversal distractor verb when both were controlled for context-based pre-activation. We controlled for the amount of pre-activation based on content by taking low-cloze responses that were either role-appropriate or role-reversed given the context, and measured whether participants produced the distractors to the same extent.

We used a set of responses provided in Experiment 3-1 to create another distractor type for Experiment 3-2; in addition to the target and control distractors, we added a 'weak distractor' condition. Weak distractors were low-cloze, role-appropriate responses that were produced by participants given the reversed context in Experiment 3-1. For each item, we used the response

that had a similar cloze probability as the target distractor in the reversed context, i.e., role-reversal verb. By comparing the target and weak distractors, which were matched in context-based cloze probability, the only difference was whether the low-cloze response was role-appropriate ('weak' distractor) or role-inappropriate ('target' distractor) given the reversed context. If post-confrontation processes were completely role-insensitive, we should see a similar or larger rate of target distractor responses (i.e., role-reversal errors) relative to weak distractor responses given the reversed context. If post-confrontation processes were role-sensitive, we should see a greater rate of weak distractor responses than target distractor responses in the reversed context.

3.3.1 Method

3.3.1.1 Participants

Forty native English speakers recruited on Amazon Mechanical Turk participated in Experiment 3-2. None of the participants had participated in Experiment 3-1. One participant was subsequently excluded due to poor performance.

3.3.1.2 Materials

The same 60 pairs of sentence fragments, each with canonical and reversed contexts, in Experiment 3-1 were used in Experiment 3-2. The same target and control distractors were used. In addition, Experiment 3-2 included a weak distractor condition. The goal was to select role-appropriate but weakly predicted verbs given the reversed context. The weak distractors were selected based on cloze responses produced in Experiment 3-1. For each item, we examined all responses produced given the reversed context in the control distractor condition and selected a response that met two conditions: i) it was a role-appropriate verb given the reversed context, and ii) it had approximately the same cloze probability as the target distractor verb in the reversed context, i.e., role-reversal. The mean cloze probability of the selected weak distractors and the original target distractors,

given the reversed context with the control distractor in Experiment 3-1, was comparable [$t(59) = -.64$, $p = .52$]. We took the matched cloze probability to represent similar activation of the weak and target distractor verbs based on the reversed context.

Each participant saw a total of 240 sentence fragments (120 critical items, 120 filler items), to which they produced cloze completions. For the critical items, 60 were presented in the canonical context and the other 60 were presented in the reversed context. In each condition, 20 were presented with target distractors, 20 were presented with control distractors, and the remaining 20 were presented with weak distractors.

3.3.1.3 Procedure

The procedure was identical to Experiment 3-1.

3.3.1.4 Analysis

The analysis was identical to Experiment 3-1. The statistical analyses for target and weak verb response rates were compared to the baseline control condition. A total of 4505 responses were analyzed. A simple difference contrast coding was used for the statistical comparisons (target: −0.5, weak: 0, control: 0.5).

3.3.2 Results

3.3.2.1 Distribution of non-distractor responses

Figure 15 shows the distribution of non-distractor responses.

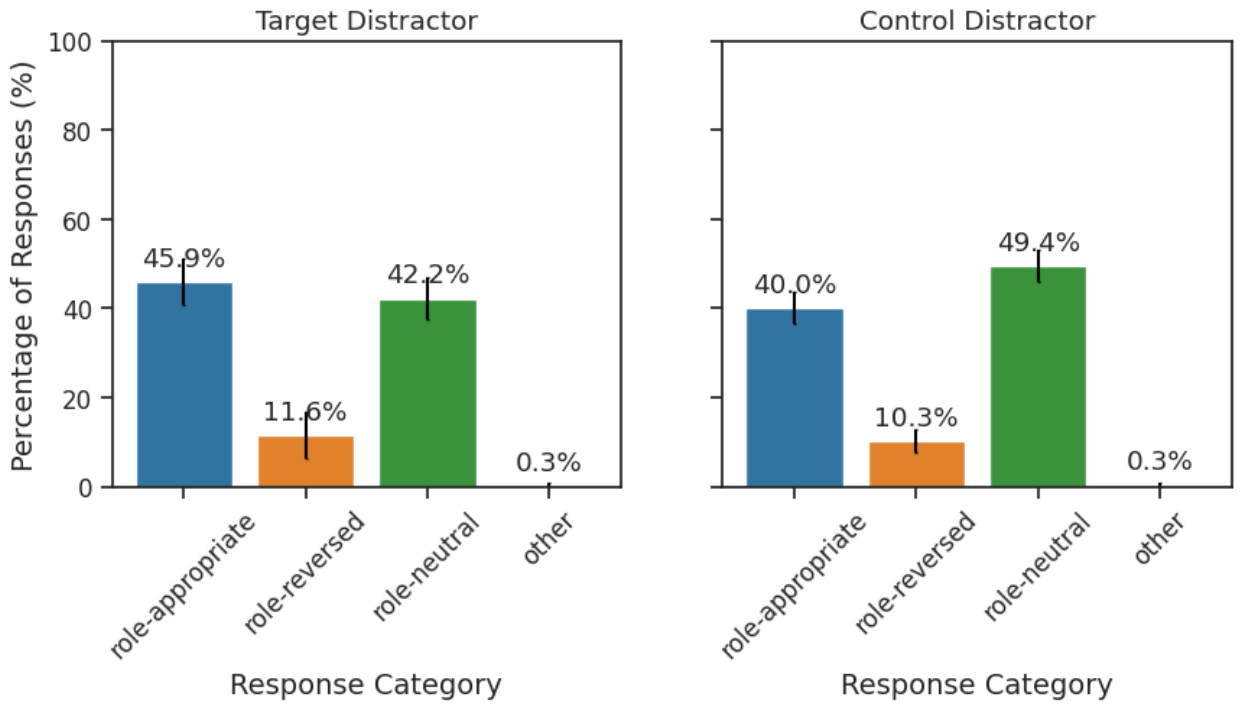


Figure 15: Distribution of responses excluding distractor responses in Experiment 3-2.

Responses were mostly role-appropriate or role-neutral and role-reversal errors were rare, replicating the pattern found in Experiment 3-1 and in previous speeded cloze studies without the distractor paradigm (Nakamura et al., 2024; Lee & Phillips, under review). This reinforces the observation that speeded cloze patterns are typically highly role-sensitive, as well as the alternative responses that participants produce in the distractor paradigm.

3.3.2.2 Cloze percentages and RTs of distractor responses

Figure 16 represents target and weak verbs produced in each condition. With control distractors, target and weak verbs were rarely produced in the reversed context, with comparable cloze percentages, replicating the cloze percentages observed in Experiment 3-1. Crucially, the rate of both target and weak verb responses increased significantly when they were seen as distractors.

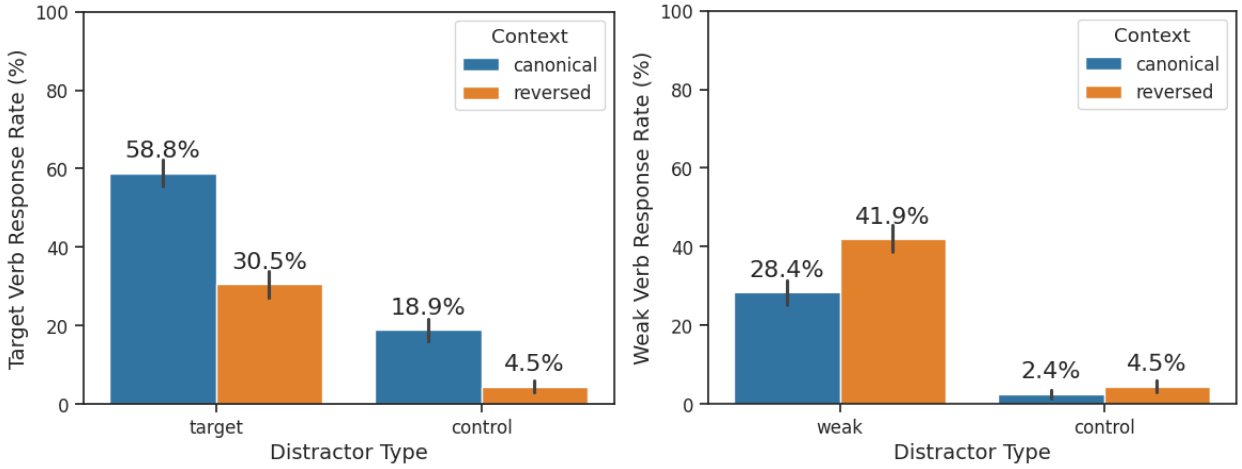


Figure 16: Experiment 3-2 cloze percentages of target verbs (e.g., *served*) (left) and weak verbs (e.g., *hated*) (right) in each condition. Error bars indicate 95% confidence intervals.

The model for target verb response rates indicated strong evidence for an effect of distractor type (β = −3.06, 95% CrI [−3.69, −2.40]), where more target verbs were produced with target distractors than control distractors. The model also showed evidence for an effect of context (β = −1.98, 95% CrI [−2.47, −1.51]), where target verb responses were reliably more frequent in the canonical context than in the reversed context. There was also evidence for an interaction between context and distractor type (β = −1.02, 95% CrI [−1.85, −0.22]). Nested models splitting the data into canonical and reversed contexts indicated a greater posterior mean of the effect of distractor type in the reversed context (β = −3.56, 95% CrI [−4.49, −2.62]) than in the canonical context (β = −2.41, 95% CrI [−3.04, −1.76]), replicating the pattern found in Experiment 3-1 (i.e., greater increase in target verb production in the role-reversed context than the role-appropriate context).

The model for weak verb response rates indicated an effect of distractor type (β = −3.95, 95% CrI [−4.75, −3.12]), where weak verbs were more frequently produced with weak distractors than control distractors. The model also showed evidence for an effect of context (β = 1.02, 95% CrI [0.62, 1.43]), where more weak verbs were produced in the reversed context than in the

canonical context. There was no strong evidence for an interaction between context and distractor type (β = 0.10, 95% CrI [−0.74, 0.94).

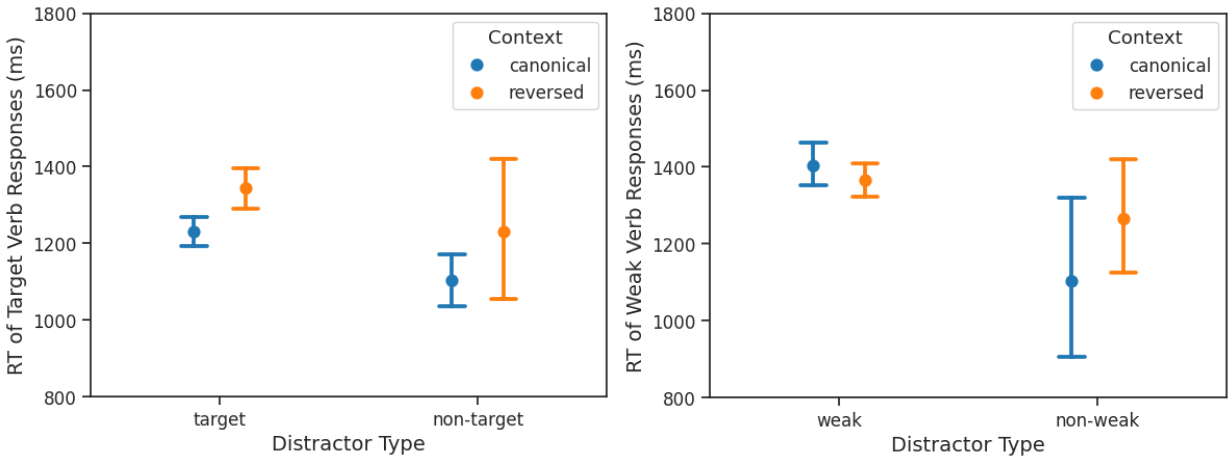


Figure 17: Experiment 3-2 cloze RTs of target verb responses (e.g., served) following target and non-target (weak and control) distractor presentation (left) and weak verb responses (e.g., hated) following weak and non-weak (target and control) distractor presentation (right). Error bars indicate 95% confidence intervals.

The RTs of target and weak verbs also showed significant differences across conditions (Figure 17). The model for target verb responses in the canonical and reversed contexts, comparing between target and control distractor presentation, was consistent with an effect of context (β = 0.14, 95% CrI [0.06, 0.22]), where target verbs were produced with slower RTs in the reversed context than in the canonical context. There was also evidence for an effect of distractor type (β = −0.13, 95% CrI [−0.23, −0.05]), where target verbs were produced with slower RTs following target distractors than control distractors. There was no strong evidence for an interaction between context and distractor type (β = 0.02, 95% CrI [−0.14, 0.18]).

The model for weak verb responses in the canonical and reversed contexts, comparing between weak and control distractor presentation, was consistent with an effect of distractor type (β = −0.22, 95% CrI [−0.36, −0.07]), where weak verbs were produced with slower RTs following weak distractors than control distractors. There was no strong evidence for an effect of context (β

= 0.08, 95% CrI [−0.05, 0.20]) or an interaction between context and distractor type (β = 0.22, 95% CrI [−0.05, 0.47]).

We also compared the rate and RTs of target and weak distractor responses when they were presented as distractors, i.e., when the presented distractor was produced as the cloze response (Figure 18).

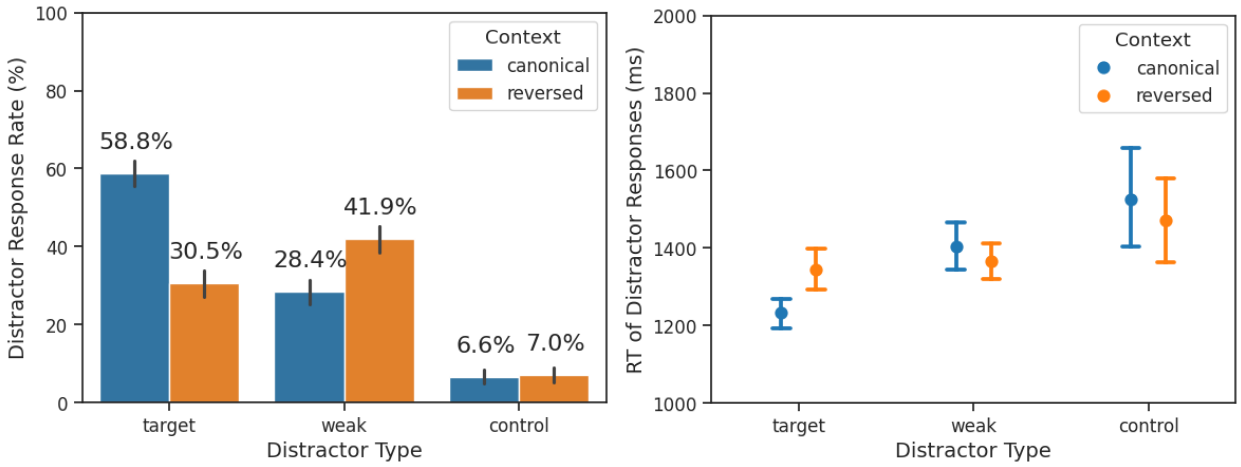


Figure 18: Experiment 3-2 cloze percentages of target, weak, and control verbs when seen as distractors (left) and their mean RTs (right). Error bars indicate 95% confidence intervals.

The model for target and weak distractor response rates indicated strong evidence for an effect of distractor type (β = −0.55, 95% CrI [−0.88, −0.23]), where participants overall produced more target distractors than weak distractors. No strong evidence for an effect of context (β = −0.19, 95% CrI [−0.41, 0.04]) was found, but there was evidence for an interaction between distractor type and context (β = 2.59, 95% CrI [1.96, 3.23]). Nested models splitting the data into canonical and reversed contexts yielded strong evidence for an effect of distractor type in both contexts, but with a greater posterior mean for the canonical context (β = −1.97, 95% CrI [−2.49, −1.48]) than for the reversed context (β = 0.82, 95% CrI [0.40, 1.26]), indicating a larger difference between target and weak distractor responses rates in the canonical context than in the reversed context.

The model for control distractor response rates with control distractor presentation in canonical and reversed contexts indicated no strong evidence for an effect of context (β = −0.09, 95% CrI [−0.78, 0.54]), indicating that the control distractor was produced to similar extents in the two contexts, when presented as a distractor.

The model for target and weak verb RTs following distractor presentation indicated strong evidence for an effect of distractor type (β = −0.08, 95% CrI [0.03, 0.13]), where target distractor responses were faster than weak distractor responses. There was also strong evidence for an effect of context (β = 0.04, 95% CrI [0.01, 0.08]), where distractors were produced faster in canonical than in reversed contexts. There was also evidence for an interaction between distractor type and context (β = −0.17, 95% CrI [−0.26, −0.08]).

The model for control distractor response RTs with control distractor presentation in canonical and reversed contexts indicated no strong evidence for an effect of context (β = −0.04, 95% CrI [−0.16, 0.08]), indicating that the control distractor was produced with similar speeds in the two contexts, when produced after it was seen as a distractor.

3.3.3 Discussion

The main aim of Experiment 3-2 was to replicate the effect of distractor confrontation observed in Experiment 3-1, in addition to clarifying whether the confrontation of an equally low-cloze but role-appropriate distractor would lead to the same amount of increase in response rates as confronting a role-reversal distractor. The results from Experiment 3-2 overall replicated the patterns observed in Experiment 3-1, with regard to target verb response rates and RTs. The distractor response patterns additionally indicated that processes that occur after the confrontation of the distractor are not entirely role-insensitive; participants were more likely to produce a role-

appropriate distractor than a role-reversal distractor, when both distractors were matched in context-based cloze probability.

The non-distractor response distribution indicated that participants successfully produced speeded cloze responses that fit the preceding argument roles, in trials where they avoided producing the presented distractor. The proportion of role-reversal errors was rare, as observed in Experiment 3-1. These results indicate that adding the weak distractor condition did not affect the overall role-sensitivity in the speeded cloze interference paradigm and that confronting a role-appropriate, role-inappropriate, or unrelated distractor does not significantly change the alternative responses speakers produce in a speeded cloze paradigm.

The target distractor response patterns in Experiment 3-2 replicated the pattern observed in Experiment 3-1, where participants produced target verbs more in the canonical, role-appropriate context than in the role-reversed context, with faster speed. The cloze percentage increased significantly when the target verbs were seen as a sentence continuation, as a distractor, more in the reversed context than in the canonical context. The large increase in role-reversal errors after confrontation suggests that even when participants themselves did not strongly generate a role-reversal continuation based on the context, they were highly tempted to produce a role-reversal continuation when they saw the lure word presented in the input.

A novel observation in Experiment 3-2, which sheds light on the role-sensitivity of post-confrontation processes, was that the increase in distractor verb responses after confrontation were greater for role-appropriate verbs than role-reversal verbs, when both were equally low-cloze continuations. That is, participants were more tempted to produce a low-cloze distractor when it was role-appropriate than when it was role-inappropriate. This indicates that during the decision process immediately before articulation of a response, speakers have the role-sensitivity that

prevents them from producing role-reversal distractors more than low-cloze role-appropriate distractors.

As a follow-up to Experiment 3-1 and 2, we aimed to further inspect what are the factors that lead to role-reversal errors. In the following section, we carried out item-wise analyses, in order to further examine which items caused more role-reversal errors than other items.

3.4 Item-wise variability

The large increase in role-reversal errors with the distractor paradigm made it possible to examine item-wise variability across experimental items. Figure 19 presents the by-item rate of target verb responses, with the target distractor presentation and with the control distractor presentation. The plot shows that there is variability across items, where for some items, seeing the distractor increased the production rate of the distractor drastically relative to when a control distractor was presented, while for some items, the increase was smaller.

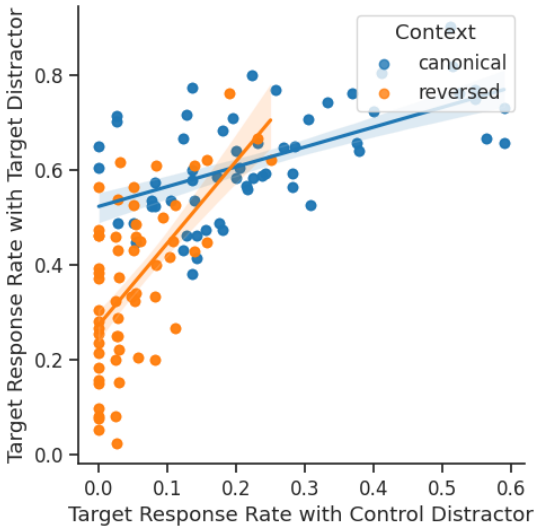


Figure 19: Relationship between rate of target distractor responses with target distractor presentation and with control distractor presentation. Each dot represents one experimental item.

For the item-wise analyses, we explored the key factors that could capture the variation in the increase of role-reversal responses observed across items. As indicated below, for each analysis,

we calculated the corresponding value for each item and examined its correlation with the size of the increase in the target verb response rate (i.e., increase in target distractor response rate = target distractor response rate with target distractor presentation *minus* target distractor response rate with control distractor presentation).

First, we examined whether the increase in target distractor response rate correlated with how often the target verb was produced in the canonical context (with a control distractor). That is, if the target verb was a high-cloze response given the canonical context, we expected that to be associated with a greater increase in the target verb's response rate when that verb was seen as a distractor. This hypothesis was not borne out. Figure 20 shows that in the reversed context, there was no evidence for a relationship between the target verb's cloze probability (in the canonical context) and the increase in target distractor response rate with distractor confrontation [$r(58) = .10$, $p = .47$]. For the canonical context, the higher the target verb's cloze probability in the control distractor condition, the smaller the increase was with target distractor presentation [$r(58) = -.70$, $p < .001$]. This is expected, since the items that led to a high-cloze target response without distractor confrontation would be less affected by the target distractor presentation, whereas the items that did not elicit a high-cloze target response would benefit more from the target distractor presentation.

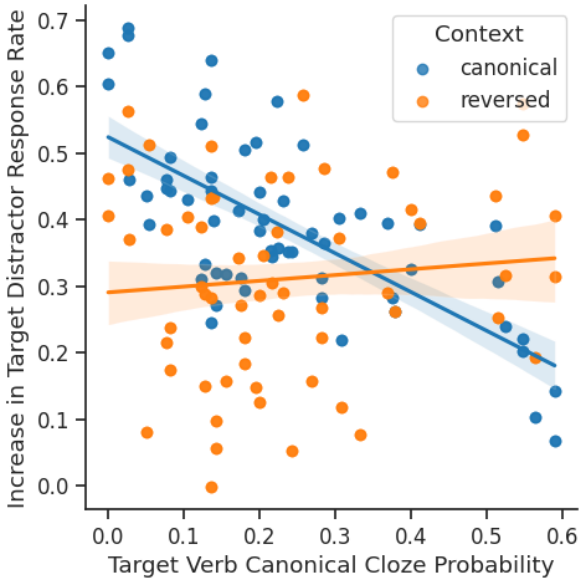


Figure 20: Relationship between by-item increase in target distractor response rate after confrontation (target distractor presentation minus control distractor presentation) and target verb cloze probability given the canonical context with control distractor presentation. Each dot represents one experimental item.

In the next analysis, we examined whether properties of the reversed context affected the rate of increase in target verb response rates. The item-wise analyses revealed that both the modal cloze probability of the reversed context and the mean response speed of the reversed context modulated the increase in target distractor response rates. Figure 21 (left) shows that for the reversed context, the higher the reversed context modal cloze probability, the smaller the increase was in target distractor response rate, i.e., role-reversal error rate [$r(58) = -.59$, $p < .001$]. For the canonical context, there was no relationship between reversed context modal cloze probability and the increase in target distractor response rate [$r(58) = -.17$, $p = .19$]. Figure 21 (right) shows that for the reversed context, the slower the mean response speed, the greater the increase was in target distractor response rate, i.e., role-reversal error rate [$r(58) = .57$, $p < .001$]. For the canonical context, no relationship was observed between the reversed context mean response speed and increase in target distractor response rate [$r(58) = -.04$, $p = .75$].

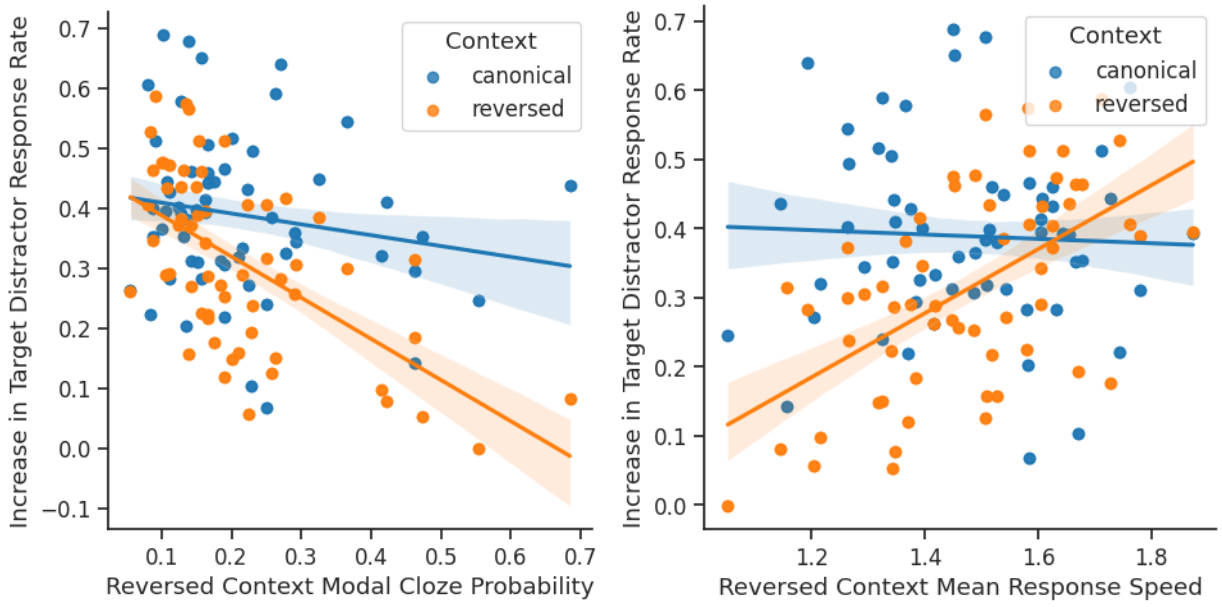


Figure 21: Relationship between the increase in target distractor response rate after confrontation (target distractor presentation minus control distractor presentation) and reversed context modal cloze probability (left) and reversed context mean response speed (right). Each dot represents one experimental item.

The key takeaway from the item-wise analyses was that the less constraining the context was, i.e., when the context did not yield a strong modal response or did not quickly generate any candidate, the more likely it elicited a role-reversal distractor response. That is, the more difficult it was to generate a sentence continuation based on the given context, the more likely participants were to produce a role-inappropriate continuation that was presented as a distractor. Alternatively, there was no significant relationship between the strength of the verb given the canonical context and how more likely the verb was produced in the reversed context when seen as a distractor, i.e., increase in role-reversal error rate.

3.5 General Discussion

In the present study, we set out to investigate whether the vulnerability to role-reversals typically observed in comprehension arises after, rather than before, the verb is encountered in a sentence context. Previous studies using comprehension experimental paradigms have mostly blamed faulty

context-based generation processes for comprehenders initially failing to differentiate between verbs presented in role-appropriate and role-reversed contexts (Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Nakamura et al., 2024). This stands in contrast to results from studies using production paradigms, which indicate that speakers can rapidly use argument role information to produce role-appropriate sentence continuations (Chow et al., 2015; Nakamura et al., 2024). In order to examine the extent to which the vulnerability in comprehension measures arises from processes that occur once the word is encountered in the context, which does not occur in production settings, we designed a novel speeded cloze interference paradigm that involved presenting role-appropriate, role-inappropriate, and unrelated distractors immediately following sentence contexts, just before participants produced their own sentence continuations. The results revealed that seeing a role-reversal verb as a distractor dramatically increased the probability of the role-reversal verb being produced as the speeded cloze continuation, nearly as much as the increase in producing a (weakly predicted) role-appropriate distractor, highlighting the vulnerability to role-reversal errors after the verb is encountered it in the input.

The majority of responses that participants produced as alternatives to explicitly presented distractors showed evidence of rapid role-sensitivity in generating predictions for upcoming verbs. The experiments in the present study involved presenting target, control, or weak distractors following canonical or reversed sentence contexts. The target distractor was a predictable continuation given the canonical context and reversing the preceding argument roles made it a low-cloze, unpredictable sentence continuation given the reversed context. The control distractor was an implausible continuation that was unrelated to either context. The weak distractor added in Experiment 3-2 was an equally low-cloze response as the role-reversal target verb given the

reversed context, but role-appropriate. Speeded cloze responses in the control distractor condition, in both Experiments 3-1 and 3-2, replicated the pattern typically observed in speeded cloze production studies (Chow et al., 2015; Nakamura et al., 2024); the target verb was produced at a rate of about 25% in the canonical context and only about 5% in the reversed context, with faster onset latencies in the canonical context than the reversed context, indicating role-sensitive generation of responses. This indicates that in trials where participants avoided producing the presented unrelated control distractor, their responses were role-sensitive. Moreover, the overall distribution of responses for trials where participants avoided producing any presented distractor also indicated role-sensitivity, in both Experiments 3-1 and 3-2. The majority of responses were role-appropriate (either specific to the given role context or could equally fit either the canonical or reversed context) and only 10% of responses were role-reversal errors. These results indicate that in trials where participants avoided the presented distractors, their responses showed clear evidence of role-sensitivity; they actively used argument role information provided in the preceding context in order to produce their speeded cloze continuations.

The response patterns following target distractors in Experiments 3-1 and 3-2 revealed a dramatic interference effect of distractor confrontation. Participants were more likely to produce a role-reversal error, when explicitly presented with the role-reversal verb as a distractor. Distractor presentation increased the role-reversal error rate up to 30-40%, a significant boost from the 5% typically observed in speeded cloze studies (e.g., Nakamura et al., 2024) and in the current study in trials where participants saw the control distractors. The significant increase in the rate of role-reversal verbs after confronting them as distractors indicate that processes that occur after the verb is encountered in a sentence context are less role-sensitive compared to processes that occur without confrontation of the verb.

However, the results for the weak distractor condition in Experiment 3-2 highlight the fact that post-confrontation processes are not completely blind to argument roles. The weak distractors, which were matched in cloze probability with the target distractors in the reversed context, elicited more distractor responses than the target distractors, indicating that participants were more tempted to produce a role-appropriate distractor than a role-inappropriate distractor when both were equally weakly generated by the context. This suggests that at the point of encountering the distractor as a candidate continuation, participants used the argument role information provided in the context in order to consider the unpredicted but role-appropriate verb as a more appropriate response than a role-reversal verb.

The findings combined challenge the strict version of the view that the susceptibility to role-reversal anomalies arise from faulty context-based prediction processes. Our results indicate that even in an originally role-sensitive measure, i.e., speeded cloze production, it is possible to observe significantly reduced performance in role-sensitive prediction when role-reversals are explicitly confronted in the input and role-sensitivity is measured afterward, as in typical comprehension paradigms. The results present a shift in focus on the vulnerability of the processes that occur after a word is encountered in a sentence context, rather than putting the blame solely on faulty content-based generation processes, such as association-based predictive processes that initially ignore argument roles (Chow et al., 2016; Liao et al., 2022; Nakamura et al., 2024), or semantic cues dominating initial processing relative to syntactic cues (Kuperberg et al., 2016; Rabovsky et al., 2018; Stone et al., 2024), or heuristics-based processing based on the likelihoods of events (Li & Ettinger, 2023). It is possible that context-based generation processes are highly role-sensitive, as reflected in speeded cloze patterns, while what is measured in comprehension studies at least partially reflects processes that occur once the verb is confronted in the input. The

speeded cloze interference paradigm offers a way to pinpoint which parts of the underlying processes are more vulnerable than others.

The item-wise analyses revealed further insights into the interaction between context-based, proactive processes and post-confrontation, reactive processes. The items that either yielded a high-cloze modal response or quickly generated alternative responses were more robust to role-reversal distractor presentation. Participants were less likely to produce the presented role-reversal verb in cases where they already had a strong alternative that was readily available. Moreover, how likely the presented role-reversal verb was in its role-appropriate and canonical context did not show a significant relationship with role-reversal distractor response rate. In other words, whether or not the role-versal was a highly predictable verb given the canonical context did not significantly contribute to the characterization of the items that led to greater role-reversal errors. Although these patterns were observed in post-hoc analyses, they converge on the point that confronting a role-reversal verb at the point when no alternative is readily available increases the degree of susceptibility to role-reversal anomalies. This reinforces the idea that rather than context-driven proactive processes, the susceptibility to role-reversal anomalies could also partly arise from reactive processes that occur once the verb is encountered, particularly when the generation process itself does not quickly yield good candidates.

The results of the item-wise analyses in the current study corroborates the patterns observed across different prior works on role-reversal anomalies. First, previous studies have shown that extending the time between the presentation of argument role information and the verb yields a significant N400 effect, i.e., role-sensitivity in comprehension (e.g., Chow et al., 2018; Nakamura et al., 2024). This has been taken to indicate that the delay in presenting the verb provides comprehenders with more time to use the given argument role information to constrain their

predictions to role-appropriate verbs. The insights from our item-wise analyses offer the possibility that this time is used to generate alternative, role-appropriate candidates. If it is the availability of role-appropriate responses that helps rescue one from considering a role-reversal verb as a good sentence continuation, then it is possible that comprehenders can benefit from additional time to generate appropriate candidates which prevents them from considering the role-reversal verb, when it is confronted in the input. Second, a previous study reported that when the canonical and reversed contexts both yield a predictable continuation (i.e., they are both similarly constraining contexts), role-reversal anomalies elicit the N400 effect (Ehrenhofer et al., 2019). This finding, in combination with the implications from our item-wise analyses, indicates that the lack of role-sensitivity observed in prior work might be at least partially due to the lack of readily available role-appropriate candidates given the reversed sentence contexts. Then, even if context-driven generation processes are role-sensitive, the empirical evidence could suggest otherwise, because responses are measured only after the role-reversal verbs are confronted in the absence of good alternatives.

A question is what are the underlying mechanisms that lead to the production of role-reversal verbs when they are presented as distractors in the speeded cloze interference paradigm and how it relates to the N400 patterns observed in EEG studies. Based on the empirical results and the insights gained from the item-wise analyses, we propose that while a generation mechanism utilizes argument role information in the context and quickly initiates a search for role-appropriate candidates, until it fully generates one or more role-appropriate candidate(s), the encoding of the argument role context is in a less stable or less structured state. In this state, the system becomes more vulnerable to explicitly imposed candidates that have many shared features with the preceding context, including role-independent argument information. This would lead to

the temptation of producing an explicitly presented role-reversal verb as a sentence continuation or eliciting a reduced N400 response to the role-reversal verb. Under this hypothesis, the shield against the tempting lures may be created either through a shift in the representation of the context into more concrete representations to afford action (Federmeier, 2022: *connecting* to *considering*) or through faster filtering of candidates that do not fit the given contextual information (Nakamura et al., 2024). The faster this shield is created, the more effectively role-appropriate and role-inappropriate verbs are evaluated when they are encountered in the input. Since different sentence contexts embed different information that can be used to quickly generate strong candidates, this could also produce the variation observed across the experimental items in the vulnerability to role-reversal distractors in the present study.

While future work is needed to clarify the exact underlying processes at the mechanistic level, our results indicate that there is a significant effect of confronting words in context which are only partially consistent with preceding contextual information. While previous accounts have mostly focused on the limitations of context-driven predictive processes in explaining the role-reversal phenomenon, our findings offer a new perspective in examining the vulnerability that arises from reactive processes that occur once an explicitly selected word is encountered as a sentence continuation.

3.6 Conclusion

In this study, we examined how explicitly encountering a word in a sentence context influences its cloze probability in a speeded cloze task. Using a novel speeded cloze interference paradigm, participants were presented with sentences word-by-word, followed by a distractor that was either contextually appropriate or inappropriate. Participants could then choose to produce the distractor

word or an alternative response. Our findings indicate that exposure to a distractor increased the likelihood of selecting that distractor verb, with a stronger effect in role-reversed contexts than in role-appropriate contexts. The increase in role-reversal errors was yet smaller than the increase in equally weakly predicted but role-appropriate distractor responses. These results, in combination with item-wise analyses, suggest that when individuals can quickly commit to a likely candidate, they are less susceptible to role-reversal errors. In contrast, lower confidence increases susceptibility to misleading lures, highlighting the dynamic interaction between context-driven predictive processes and reactive processes that occur after a word is encountered in a context during real-time sentence processing.

Chapter 4: A competitive generation mechanism where speed matters

4.1 Introduction

A large body of work in psycholinguistics has shown that adult comprehenders actively anticipate upcoming input during real-time sentence comprehension (Altmann & Mirković, 2009; Van Petten & Luka, 2012; Dell & Chang, 2014; Huettig, 2015; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2013; Pickering & Gambi, 2018; Federmeier, 2007, 2022). Supporting evidence comes from various measures, including early looks toward targets in visual world eye-tracking (Altmann & Kamide, 1999; Kamide et al., 2003), facilitated reading times in eye-tracking while reading or self-paced reading paradigms (Ehrlich & Rayner, 1981; for a review see Staub, 2015), and brain responses that are sensitive to the predictability of the target input (e.g., Kutas & Federmeier, 2000; DeLong et al., 2005). While the exact mechanisms underlying how expectations are formed are debated, it has been argued that adult comprehenders generate multiple possible candidates for upcoming words in parallel (Roland et al., 2012; Staub et al., 2015; Frisson et al., 2017; Ness & Meltzer-Asscher, 2021), instead of limiting predictions to a single most highly probable candidate. The parallel activation of multiple candidates has been a mechanism proposed to explain key empirical findings, including systematic patterns in how quickly people produce responses in a speeded cloze paradigm (Staub et al., 2015).

Studies have shown that children exhibit anticipatory behavior similar to adults during real-time sentence comprehension (Borovsky et al., 2012; Nation et al., 2003; Mani & Huettig, 2012), but whether the underlying predictive mechanism also involves activating multiple lexical candidates in parallel is less clear. The majority of evidence supporting children's ability to form expectations for upcoming input comes from the visual world paradigm, where increased looks to

a target picture before it is explicitly mentioned in the linguistic input is taken to reflect anticipatory behavior. Despite its many benefits, the method has limitations in probing the range of hypotheses children consider during prediction, or whether they consider multiple candidates simultaneously, and if so, whether they generate similar sets of candidates to adults. In the present study, we examined whether children share the same underlying predictive mechanism as adults, specifically whether they exhibit adult-like patterns in a speeded cloze paradigm which reflect the dynamics of competitive activation of multiple candidates.

4.1.1 Underlying mechanisms in prediction: Sampling from a probability distribution, or competition among multiple candidates?

The term *prediction* has been used in the psycholinguistics literature to refer to different processes with different assumptions about the underlying mechanisms. The present study views prediction in the form of context-driven pre-activation (DeLong et al., 2005; Kutas et al., 2011; Staub et al., 2015; Federmeier, 2022). Adult comprehenders can use contextual information in combination with their prior knowledge of the world and linguistic constraints to probe long-term memory and activate relevant items that will help them to process upcoming linguistic input. This kind of pre-activation can occur at multiple levels, including semantic features (Federmeier & Kutas, 1999) and lower-level phonological forms (Mani & Plunkett, 2010; Gambi et al., 2018). Here, we focus on the activation of lexical candidates that are considered to be a good 'fit' to the preceding context. Researchers have often measured this goodness-of-fit by using the cloze probability of a word given a context, where cloze probability is calculated by how frequently a word is produced as a continuation of a given sentence fragment within a group of people. Higher cloze probability is taken to indicate a better fit of the word to the context, relative to other potential candidates that have a lower cloze probability.

While cloze probability has been widely used to represent how predictable a particular word is based on its preceding context, exactly how it maps onto the actual strength of a candidate in relation to other potential candidates is less clear. Consider two possibilities that have been proposed in the literature, regarding the connection between cloze probability and the underlying mechanism that leads to responses with different cloze probabilities for a given context. One view is that cloze probability is derived from a process where each individual samples from a subjective probability distribution (e.g., Smith & Levy, 2011). Given world and linguistic constraints, people have prior distributions of likely continuations for a context, and in a cloze task, they sample from this subjective probability distribution and produce a cloze response. In this way, the proportion of responses produced for a given context directly reflects the probability distribution of those candidates. Although cloze probabilities can diverge from actual corpus-based conditional probabilities, potentially due to biases in people's subjective prior distributions (Smith & Levy, 2011), the key claim of this approach is that cloze probability is a direct reflection of how good a candidate is given a context, because people simply sample from a subjective probability distribution when they produce responses in a cloze task.

A different view on cloze probability is offered by Staub et al. (2015), where it is claimed that cloze probability is the output of a competitive race process: multiple candidates gain activation based on context and the first candidate to reach an activation threshold gets produced. Unlike the sampling account, Staub et al.'s (2015) race model does not view cloze probability as a direct read-out of how likely a word is given a context. Different candidates accumulate activation and race toward the threshold with different speeds, and the primary relationship is between a candidate's strength and the speed of activation. Under this theory, the time it takes for a response to get produced, or the cloze reaction time (RT), is a more direct reflection of goodness-

of-fit than cloze probability, although it is not a direct reflection, since RTs are only observed for winners, i.e., responses that are produced. The key point is that the speeds of candidates are fed into the race process, and cloze probabilities reflect win proportions. In Staub et al.'s (2015) race simulations, the strength of each candidate in a given context is operationalized through a mean finishing time distribution, where this distribution plus some random noise together yield slightly different outcomes on multiple trials of the same race. A candidate's own speed and the speed of its competitors jointly determine the probability of how often the candidate wins the race, represented by cloze probability. Staub et al.'s (2015) race model does not make assumptions about direct influence among competitors on their activation profiles, e.g., via lateral inhibition. In this model, the effect of different competitors on each other is limited to determining whether or not they are winners. We follow this assumption here, but see Nakamura (2023) for evidence for lateral inhibition effects in a speeded cloze paradigms.

The race model offers a direct explanation for systematic timing patterns found in adults' speeded cloze responses which are not straightforwardly explained by the sampling account. In Staub et al.'s (2015) speeded cloze experiments, participants read sentence fragments presented in rapid serial visual presentation (RSVP) and produced cloze responses under time pressure. The authors discovered two key patterns in response latencies. The first was that high-cloze responses were produced, on average, more quickly than low-cloze responses. The race model naturally predicts this relationship between cloze probability and response times, where faster candidates have a better chance of winning the race and getting produced than candidates that are slow. Therefore, more frequently winning candidates should have faster finishing times. The sampling account proposed by Smith & Levy (2011) does not make any predictions regarding response latencies, as the production of a cloze response is a result of a one-time sampling from a probability

distribution. However, since the original claim was proposed to explain offline cloze probabilities, the account could be extended to explain speeded cloze response latency patterns. The relationship between the subjective probability distribution and cloze probability could be used to assume a similar relationship between subjective probability and response latencies. If there is a core notion of strength of fit between a context and a candidate, this could directly impact both the probability of sampling and the speed of the response. In this way, the account could capture the fact that high-cloze responses are faster than low-cloze responses.

The second key finding in adults' speeded cloze response latencies observed by Staub et al. (2015), however, more clearly favors the race model over the sampling account. The authors found that responses are generally faster when produced in a highly constraining context than in a low constraining context, where context constraint was measured by the cloze probability of the modal response, i.e., the most frequent response given a context. This pattern naturally follows from a competitive race process, where candidates that are in a race with fast competitors can only reach the threshold more quickly than the competitors when they are themselves relatively fast. A candidate is only produced when it is activated faster than its competitors, and therefore, a candidate with faster competitors needs to be faster to win than a candidate with slower competitors. The relationship between contextual constraint and speeded cloze response latencies is not predicted by the sampling account, because a candidate's strength (and hence, presumably, its speed of generation) is taken to be directly reflected in its subjective probability. There is no parallel activation of multiple candidates, and hence no reason for the speed of a response to be affected by the strength of alternatives. The fact that responses are produced more quickly when their alternatives are stronger does not follow without additional assumptions from the sampling model.

Based on these empirical findings, there is strong support for the claim that adult comprehenders actively consider multiple possible candidates that fit the preceding context in parallel, and that the relative strengths of the generated candidates affect how likely a candidate gets produced as a cloze response. In the next section, we discuss the possibility of whether children also utilize a race mechanism in predicting upcoming linguistic input during real-time sentence processing.

4.1.2 Children's predictive processing

Research on children's predictive processing has mostly focused on whether or not children predict upcoming input like adults. Many studies report findings indicating that children from a very young age show predictive behaviors similar to adults during real-time comprehension. One common method used to study prediction in children is the visual world eye-tracking paradigm, where the proportion of looks toward different objects displayed on a visual scene is measured while a sentence is aurally presented. Early increased looks toward a target object before it is mentioned in the sentence is taken as evidence for prediction, or anticipation of the target before it is explicitly provided in the linguistic input. Previous studies using the visual world paradigm have found that children as young as two years old exhibit adult-like anticipatory looking patterns (Nation et al., 2003; Mani & Huettig, 2012; Borovsky et al., 2012). For example, Borovsky and colleagues (2012) aurally presented sentences like, *The pirate chases the ship*, while 3-to-10-year-old children viewed a display of four objects: a treasure box, a pirate ship, a cat, and a bone. Upon hearing *chases*, the proportion of looks toward the correct object, the pirate ship, increased in comparison to the other objects. This was taken as evidence that young children are capable of using their world knowledge about pirates and a chasing event to quickly start looking toward the object representing a likely patient given the context. Similar results were found with 2-year-olds

(Mani & Huettig, 2012) and 10-to-11-year-olds (Nation et al., 2003). Studies have also shown that children actively use syntactic structure (Gambi et al., 2016) and morpho-syntactic cues that are available in their language for prediction, such as grammatical gender or number agreement (Lew-Williams & Fernald, 2007; Lukyanenko & Fisher, 2016; Özge et al., 2019, 2022; Aumeistere et al., 2022; Brown et al., 2022; Smolík & Bláhová, 2022).

The previous findings highlight parallels between children's and adults' predictive processing primarily based on measures like eye-movements and responses to visually presented stimuli. The visual world paradigm, as described above, is a popular experimental method for probing children's predictive behavior due to its many advantages. It allows the experimenter to control the kind of context that is available to the participant. It is easily combined with auditory sentence presentation, and no reading is required. It provides moment-by-moment information about when the context starts to affect eye-movements. It also makes it possible to observe effects that occur before the target is present in the linguistic input, indicating anticipatory behavior. Predictive eye-movements can therefore tell us how efficiently children can analyze the given linguistic context and combine their analysis with prior world knowledge and awareness of the current visual scene to generate expectations for upcoming input.

Despite these benefits, the visual world paradigm, along with other comprehension measures like EEG/ERPs, has limitations in probing whether there is parallel activation of multiple candidates during prediction. To understand how comprehenders combine linguistic context with memory to generate candidates, researchers have often used anticipatory looks to targets simply as a read-out of that process. However, the method involves explicitly presenting candidates through a visual display, which makes observation of the candidates that are specifically generated based on the linguistic context challenging. To take one example, in Borovsky et al. (2012),

children looked toward an appropriate target object before it was mentioned in the linguistic input, indicating that they had anticipated the target based on the preceding context. The authors, however, also observed a greater proportion of looks toward an action-related distractor that was an inappropriate patient of the verb given the preceding agent but semantically associated with the verb. For instance, upon hearing, "*The pirate chases the,*" children initiated more looks toward the correct target, ship, but also looked toward the verb-related distractor, cat, more than the other distractors, bone and treasure box. Increased looks toward the verb-related distractor might suggest that the contextually inappropriate candidate was considered in parallel with the target appropriate candidate. However, given the design of the visual world paradigm, it is possible that the presence of inappropriate candidates on the visual display led the comprehender to consider candidates they would not otherwise generate in the absence of the visual input. The cat may act as a competitor only because it is presented as a potential target. Therefore, it is difficult to observe the candidates that children generate based on context in absence of explicitly presented candidates in a visual world paradigm. A similar concern is relevant in EEG/ERP paradigms, where responses to presented target words are measured, rather than directly probing self-generated predictions in advance of seeing the target words.

Another limitation of using comprehension-based experimental methods is that it restricts the type of data that can be used to probe whether there is parallel activation of multiple candidates. In a visual world paradigm, participants can only look at one target object at a given time, and aggregating across many individual looks does not straightforwardly show evidence for simultaneously active candidates. For example, a recent study reported that, like adults, 5-to-6-year-old children entertain multiple possibilities for upcoming words in a visual world paradigm (Sommerfeld et al., 2023). Upon hearing, "*The father eats now the…*," children directed looks

toward pictures of objects that were potential candidates that fit the semantic constraints of the preceding verb (i.e, edible objects), and they were equally distributed across multiple possible targets when presented with multiple visual objects that fit the context (pretzel, pizza, sausage, waffle). The authors concluded that children, like adults, consider multiple possibilities when predicting upcoming words based on preceding context. These findings are entirely consistent with the sampling approach (Smith & Levy, 2011), where a single response is sampled from a probability distribution on each trial. Determining whether the equal proportion of looks toward the different objects on the screen reflects parallel activation of those candidates and whether the relative strengths of candidates play a role in the process is less straightforward (see Gambi et al., 2021 for alternative ways of measuring graded predictions using the visual world paradigm).

4.2 Experiment 4

We took advantage of the specific linking hypotheses between measures and underlying cognitive mechanisms proposed by the race model (Staub et al., 2015) and used the speeded cloze paradigm to probe the predictive mechanisms used by children. One possibility is that children and adults share the same mechanism and that they generate similar candidates for a given context. In that case, we would expect the same kind of competitive dynamics to appear in children's speeded cloze RTs as are found in adults. An alternative possibility is that the underlying mechanism is shared between children and adults but that the candidates and their properties differ between the two groups. For example, children might struggle to generate multiple viable candidates on some trials, due to differences in their world knowledge, vocabulary, or memory access mechanisms. In this case, we might still find evidence for systematic timing profiles in children's cloze responses, but we might not find the same effects of competition between candidates. It is also possible that children do not use the same race mechanism as adults, and that the race mechanism develops later

in life. For example, children might sample words from a probability distribution (Smith & Levy, 2011), which does not involve competition among simultaneously generated candidates but rather a one-time sampling from a prior subjective distribution. Finally, it is possible that children's cloze behavior varies significantly across individuals, making it difficult to generalize a common mechanism based on group-based measures like cloze probability and RTs. This could be either because children do not have a shared mechanism as a group, or because they do not have the same kind of distribution of candidates which they sample from or use to generate candidates. Response patterns then may not be accurate reflections of a shared mechanism, as they would depend on each child's knowledge of linguistic and world constraints and other individual differences.

We examined whether children's speeded cloze responses exhibited the same pattern reflecting competitive dynamics as found with adults and as predicted by the race model. The study was carried out as part of a partnership with the Language Science Station at Planet Word, a language-focused museum in Washington D.C.. The Language Science Station brings research and researchers into the museum, with the aim to conduct research in a way that is mutually beneficial for researchers, students, and for museum visitors. Students receive training in science communication and conducting research via a summer course, and they staff booths at the museum to run different studies and talk with museum visitors about research. Carrying out the study in a museum setting made it possible to recruit a diverse group of participants, including school-aged children. The unique experiment setting also prompted the design of a child-friendly gamified version of the speeded cloze paradigm to fit the purpose of the research at the museum.

To preview the findings, children's and adults' response rates and RTs revealed the hallmark patterns reflecting race dynamics in both children and adults, supporting a shared mechanism in generating real-time predictions. There were, however, differences between children

and adults regarding RT evidence supporting the race mechanism. We show through a comparison of different measures of sentence constraints and race model simulations that the children's speeded cloze patterns that diverged from adults' can nevertheless be captured through a race process, where the outcomes are largely affected by the properties of simultaneously activated candidates.

4.2.1 Method

4.2.1.1 Participants

Participants were native English-speaking school-aged children (N = 60; ages 4-12, mean 9) and adults (N = 136; ages 18-59, mean 34). All participants voluntarily participated in the experiment as part of their experience at the Planet Word museum in Washington D.C., USA. Participants filled out a language background questionnaire at the beginning of the experiment which was used to screen participants' language and age background. All participants were sufficiently proficient readers to carry out the experimental task.

4.2.1.2 Materials

The experimental stimuli consisted of 40 sentence fragments with varying sentence constraints. They were created by truncating sentences with a range of high and low cloze probabilities, taken from a sentence norming study (Block & Baldwin, 2010). Table 3 presents examples of a high-cloze and low-cloze item. The 40 items were divided into two presentation lists, such that each participant saw only one of the two lists. Each list also included the same 20 filler items used for another experiment not reported here, consisting of object relative clause structures with a manipulation of word order to test the use of argument roles in verb prediction. In total, each

participant saw 40 items, 20 critical items and 20 fillers, presented in random order. The full set of stimuli can be found in the Appendix.

Table 3: Example experiment stimuli

Sentence fragment	**Example response**	**Normed cloze probability***
Water and sunshine help plants	grow	0.95
A large stone blocked the entrance to the	cave	0.36

*The normed cloze values are from Block & Baldwin (2010)

4.2.1.3 Procedure

Gamification of the speeded cloze paradigm

We designed a gamified, child-friendly version of the speeded cloze task, where the setup for the study introduced the idea of a robot as an example of an auto-complete technology. Participants were told that they would play a language science game called, 'Race the Robot,' and the goal of the game was to silently read sentence fragments and say aloud a continuation as quickly as possible. This version of the speeded cloze paradigm had the benefit of ensuring that participants, including young children, were engaged in the experiment in the less controlled museum setting. Our replication of the key adult findings from Staub et al. (2015), which used a speeded cloze paradigm in a more typical lab setting, confirms that the museum setting and the modified speeded cloze paradigm in the current experiment did not elicit different predictive behavior. Although not reported here, we also carried out a follow-up experiment using the same gamified speeded cloze paradigm and identical stimuli with adult English speakers on Amazon Mechanical Turk and replicated the main findings. Since the same results were obtained whether participants carried out the task in a quiet home setting or in the noisy museum setting, it is unlikely that the museum setting affected the results. The details of the online study can be found in Lee et al. (2023).

Task instructions were provided in verbal and written form, with three interleaved practice trials. In the practice trials, after participants produced a cloze response, a picture of a robot and an example completion of the sentence fragment were presented (Figure 22), to which participants could compare their own responses. No particular judgment was given by the experimenter about the responses that the participants produced, and no feedback was provided during the actual experiment. All participants provided verbal consent and completed a demographic survey before beginning the experiment. At the end of the experiment, the experimenter engaged in a conversation with participants about the purpose of the study and its potential relevance to everyday language use, as part of the learning experience at the museum. This was a distinctive feature of conducting research at a language museum and more broadly, promoting language science to the general public. Students received extensive coaching in how to hold these conversations, and museum visitors were highly engaged.

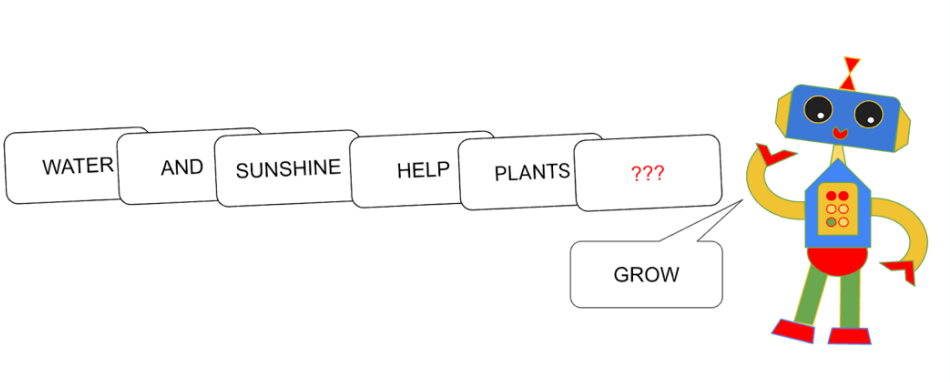


Figure 22: Illustration of the gamified speeded cloze experiment 'Race the Robot'.

The experiment was administered using PCIbex (Zehr & Schwarz, 2018) on laptop computers. Each trial began with a '+' fixation mark presented for 1000 ms, followed by a sentence fragment presented word-by-word, with a 530 ms SOA (300 ms per word with a 230 ms blank screen in between). After the final word, a '?????' sign appeared, during which participants produced a completion. The screen automatically moved onto the next trial after a three-second

time limit. Participants' responses were recorded using a directional microphone that was robust to surrounding noise in the museum. Each experiment session took about 10 minutes to complete, including consent and practice.

4.2.1.4 Analysis

4.2.1.4.1 Processing of speech data

The collected audio files were pre-processed by obtaining automatic transcriptions using Google Cloud Speech-to-Text API and by detecting automatic speech onset times using Chronset (Roux et al., 2017). The transcriptions and onset times were then manually checked and adjusted using Praat (Boersma, 2001). The review of transcriptions was carried out by the authors of this study. Codes that identified the experimental items were removed during this process, in order to avoid introducing biases in the manual review of transcriptions and onset times. Responses that were unidentifiable or that exceeded the 3-second time limit were removed from subsequent analyses.

4.2.1.4.2 Response time analyses

Cloze probabilities of responses were obtained by dividing the count of each response to a given sentence fragment by the total number of responses to that fragment, separately for children and adults. The response with the highest cloze probability for each item for each group was labeled as the item's modal response, indicating the most frequently produced response for the item. Children and adults did not always have the same modal response. Following Staub et al. (2015), and as standard practice in the literature, we initially used the cloze probability of the modal response as a measure of 'sentence constraint,' or how constraining the particular item is. However, in subsequent analyses we will show that this is not the ideal measure and propose that mean response latencies more accurately reflect sentence constraint.

Statistical tests for the main effects of interest were conducted through linear mixed-effects models of RT using the lme4 package (v1.1.35.1; Bates et al., 2015) in R (v4.3.2; R Core Team, 2023). Log-transformed RTs were used in the models, and raw RTs were used for the figures. All fixed effects were centered on the grand mean, except for the model for item constraint, which were centered on the grand means within non-modal responses. All of the models initially included a maximal random effects structure for subjects, with the subject intercept and by-subject slopes for each of the fixed effects, and were simplified until the model converged (Barr et al., 2013). Random effects for items were not included, as item effects cannot be estimated independently of the fixed effects (Staub et al., 2015).

First, a model was built to examine the effect of cloze probability on RT, which included cloze probability, group, and their interactions as fixed effects. A separate model replacing cloze probability with the modal/non-modal contrast as a fixed effect was used to examine whether RTs differed between modal and non-modal responses (high- vs. low-cloze responses) at all levels of sentence constraint. Sentence constraint was represented by the modal cloze probability of each item, where modal cloze probability refers to the cloze probability of the response that was produced most frequently for an item. Finally, a model including only the non-modal responses was constructed, with cloze probability, sentence constraint, group, and their interactions as fixed effects, to examine the effect of sentence constraint on RT and whether it was consistent across responses with different cloze probabilities. Further details of the analyses, as well as follow-up analyses, are presented in the Results section.

4.2.1.4.3 Race model simulations

Analyses of the speeded cloze data were augmented with race simulations modeling different contexts in order to show how the observed empirical patterns can be derived from a race process. Further details of the simulations are provided along with the results in the next section.

4.2.2 Results

We first present the two key patterns found in children's and adults' speeded cloze responses, followed by an interim summary relating the observed findings to the race model. Then, we show through race simulations our proposed explanations for the observed child and adult differences.

4.2.2.1 The relationship between cloze probability and RT: *"Faster candidates, more frequent wins"*

Figure 23 presents the mean RTs of children's and adults' cloze responses plotted against cloze probabilities of the responses, where each point on the plot represents an individual response-context pairing. Children had longer response latencies than adults overall, with approximately a 200 ms delay (adults' mean RT: 589 ms; children's mean RT: 830 ms), but both groups showed the same relationship between cloze probability and RT, such that responses became faster as cloze probability increased.

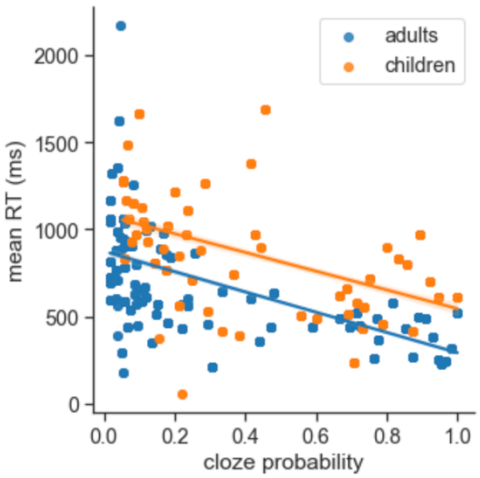


Figure 23: Relationship between cloze probability and RT of individual responses

The statistical model confirmed this pattern; there was a significant main effect of group (β = .33, *SE* = .08, $p < .0001$), indicating the overall delay in children's RTs compared to adults', a main effect of cloze probability (β = -.97, *SE* = .07, $p < .0001$), and an interaction between group and cloze probability (β = .42, *SE* = .14, $p = .004$), where both children and adults produced high-cloze responses faster than low-cloze responses, although the size of the effect was larger for adults than children.

The correlation between cloze probability and response RT, where high-cloze responses were produced faster than low-cloze responses, can be explained by a race process, as discussed in detail in the Introduction. Faster candidates have a better chance of reaching the activation threshold first and winning the race, which means they should be produced more often than slower candidates. This predicts that responses with high cloze probability should have faster RTs, a prediction that is borne out in the speeded cloze data with both children and adults. However, this pattern could also be consistent with the sampling account, assuming that the strength of fit between a context and a candidate directly impacts the speed of the response in the same way that it affects the probability of sampling.

4.2.2.2 The relationship between sentence constraint and RT: *"Faster competitors, faster wins"*

Based on the race model, we expected to find a relationship between sentence constraint and RT, where highly constraining contexts should yield faster responses, at all levels of cloze probability. A candidate in a race must be faster than its competitors to win and get produced, and a candidate must be faster to win a more competitive race in a highly constraining context than to win a race with weaker competitors, due to the competitive dynamics of the race process.

Figure 24 shows the difference between the speed of non-modal responses produced in high versus low constraint items, with responses grouped according to their cloze probabilities. When responses were matched in terms of cloze probability, the responses were produced more quickly in highly constraining contexts than in less constraining contexts. For adults, this pattern was true across all cloze probability ranges. Children showed a similar relationship between non-modal response time and contextual constraint, except in the case of low cloze responses, where contextual constraint did not affect RT (the yellow and gray lines in the children's plot converge at low cloze probabilities).

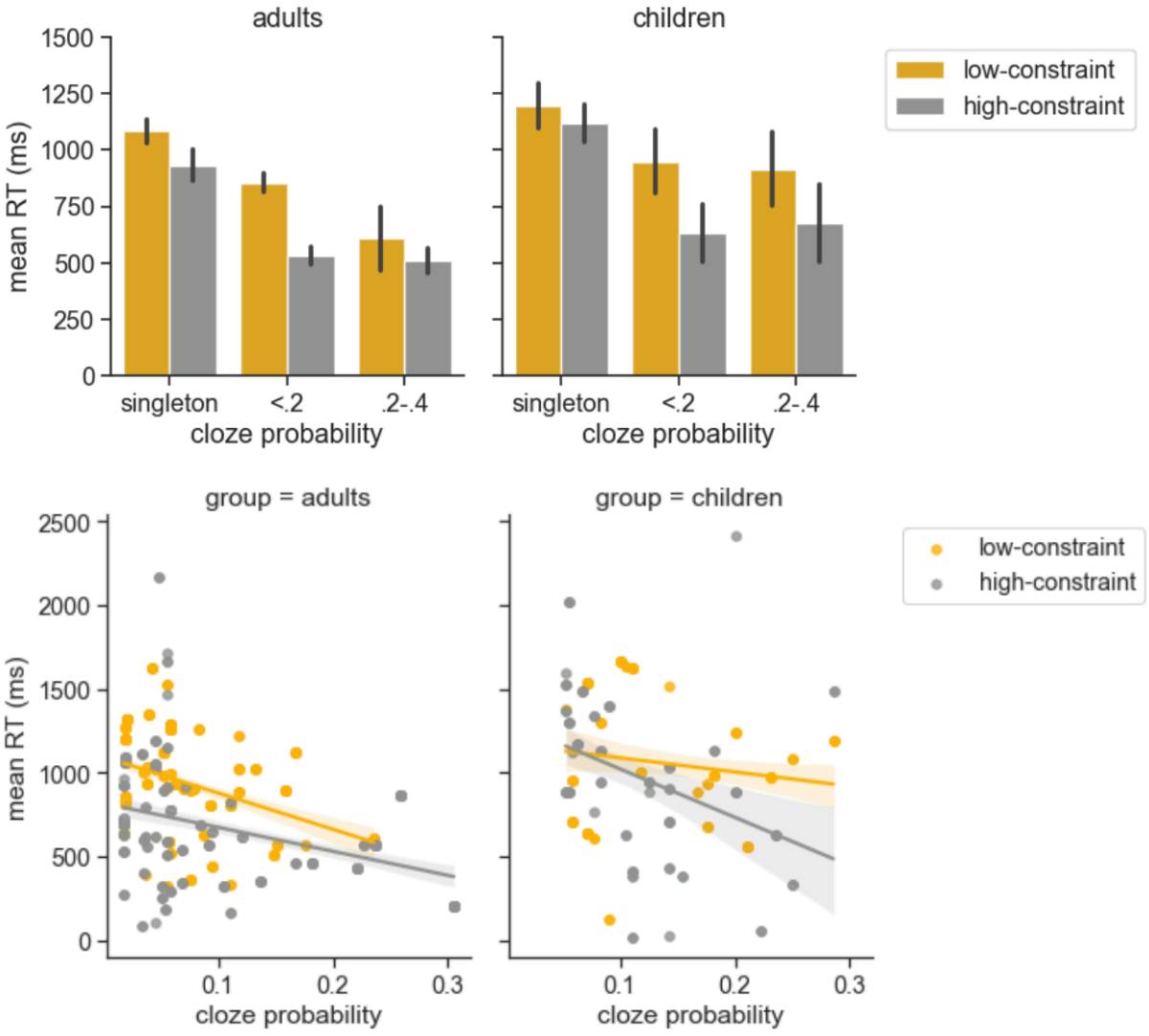


Figure 24: Relationship between sentence constraint and RT in high- and low-constraint items. Error bars represent standard error of the mean.

The statistical model revealed a main effect of group (β = .32, *SE* = .12, p = .008), a main effect of cloze probability (β = -3.69, *SE* = 0.80, $p < .0001$), a main effect of sentence constraint (β = -.65, *SE* = .17, $p < .001$), a significant interaction between cloze probability and sentence constraint (β = -7.50, *SE* = 3.21, p = .02), and a significant three-way interaction among cloze probability, sentence constraint, and group (β = -18.90, *SE* = 6.42, p = .003). In the models for each of the groups, adults' responses showed a main effect of cloze probability (β = -2.97, *SE* = .62, $p < .001$) and sentence constraint (β = -.82, *SE* = .17, $p < .001$) with no interaction between the effects (p = .47), indicating that higher cloze probability and higher sentence constraint independently contributed to faster RTs, replicating the pattern found with adults in Staub et al.'s (2015) experiments. Conversely, children's responses did not show any significant main effects but showed an interaction between cloze probability and sentence constraint (β = -16.32, *SE* = 6.51, p = .013), where sentence constraint affected high-cloze non-modal responses more than low-cloze responses. Children's low-cloze responses were relatively slow regardless of how constraining the context was, and their responses were relatively slow in less constraining contexts, regardless of cloze probability. This pattern diverges from adults, and it does not naturally follow from the competitive dynamics predicted by the race model.

In sum, the analysis of sentence constraint revealed a contrast between children and adults. Adults consistently showed faster RTs for responses produced in highly constraining contexts, at all levels of cloze probability. This naturally follows from a race process where candidates have to be faster to win a race with stronger competitors than to win a less competitive race with weaker competitors. Children did not show the same relationship between constraint and RT as adults. In

the following section, we address the divergent pattern in children by re-evaluating the measure of sentence constraint and by examining specific items where children and adults diverged the most. We show that i) the race model makes correct predictions regarding children's behavior once we use a more direct measure to classify sentence contexts based on the model's assumptions, and ii) the comparison between the different measures of sentence constraint is informative about potential differences between children's and adults' predictive processing.

4.2.2.3 Revisiting the race model predictions in relation to speeded cloze RTs

According to the race model, the average winning time of a candidate depends on the relative strength of its competitors. A race with strong competitors will generally produce fast winning times, regardless of which candidate wins the race, since a winner must be fast enough to beat the other competitors. However, a fundamental assumption of the model is that, during prediction, multiple lexical candidates are initially generated based on context, which then compete toward a threshold. If only a few candidates are initially generated, this should affect the subsequent race process, including the competitive dynamics that yield the RT patterns that are observed in speeded cloze responses.

A closer examination of children's responses for specific items shows that children may, in fact, have fewer candidates relative to adults for some contexts. An example item that elicited the most divergent patterns between children and adults is presented in (1), with the list of distinct responses and the mean RTs of all responses produced for the item, by children and adults. Modal responses are indicated in bold.

(1) ***During autumn the air is crisp and…***

	Responses	Modal cloze probability	Mean RT	Normalized* mean RT
adults	{*chilly, clean, clear,* ***cold****, cool, cream, dry, fresh, humid, light, ripe*}	0.41	695 ms	1348 ms
children	{***cold****, sunny*}	0.80	1205 ms	1784 ms
(2) ***When she got out of the car she closed the…***				
	Responses	Modal cloze probability	Mean RT	Normalized* mean RT
adults	{*car,* ***door***}	0.98	324 ms	628 ms
children	{***door****, window*}	0.88	472 ms	698 ms

*Normalized within each group, by subtracting the mean and dividing by the standard deviation.

For the context in (1), the modal response is the same in children and adults, but a diverse set of alternative responses in adults is replaced with a single alternative response in children. This suggests that children simply do not have the range of knowledge (or do not have the memory access abilities) that allows adults to generate so many different responses, and for that reason, some of the children's responses are notably slower, as reflected in the slow mean RT. These patterns contrast with the way children and adults responded to items like (2), which yielded comparable modal cloze probability, more similar mean RTs, and the same number of unique responses between the two groups. Items like (1) suggest that children and adults have different candidate profiles for some contexts, where children have fewer candidates than adults, which may lead to distinct speeded cloze RT patterns. This observation is further supported by the difference in the average number of unique responses produced for a context between the two groups, which was nine words for adults and five words for children.

The present analysis demonstrates a limitation of the standard analysis that represents sentence constraint using modal cloze probability. Doing so classifies sentence contexts like (1) and (2) into the same, high-constraint contexts for children, despite the noted differences in the candidate profiles. Moreover, it treats context (1) as more constraining for children than for adults, which is true under the standard definition of contextual constraint. However, to the extent that contextual constraint is understood to entail easy access to consistent responses, it is misleading. This motivated our choice to use an alternative measure that could more directly represent the overall strength of competitors in a given context, specifically the mean RT of all other responses produced for the context. For a particular candidate, a fast overall mean RT indicates strong competitors, and hence, a more competitive race, which should yield a faster winning time for the candidate, compared to a situation where the responses are overall slow, indicating weak competitors. We conducted an additional analysis of children's and adults' speeded cloze RTs using this alternative measure of sentence constraint and tested whether the empirical patterns naturally follow from a competitive race process, as proposed by the race model. The overall strength of competitors should affect individual response RTs due to an underlying race mechanism in prediction.

4.2.2.3.1 The relationship between sentence constraint and RT - revisited

As described above, the strength of competitors can be represented by the average of all other responses' RTs, where a fast mean RT indicates strong competitors, i.e., a more constraining context, and a slow mean RT indicates weaker competitors, i.e., a less constraining context. Based on the race model, we expected to see faster individual response RTs when competitor strength was strong and to see this relationship hold at all levels of cloze probability; a race with fast candidates will generally produce a fast winning time regardless of who the winner is.

Figure 25 shows the relationship between competitor strength and the RTs of individual responses. Unlike what was observed with modal cloze probability, the new analysis showed that at all levels of cloze probability, and in children and adults alike, responses were produced more quickly in highly constraining contexts.

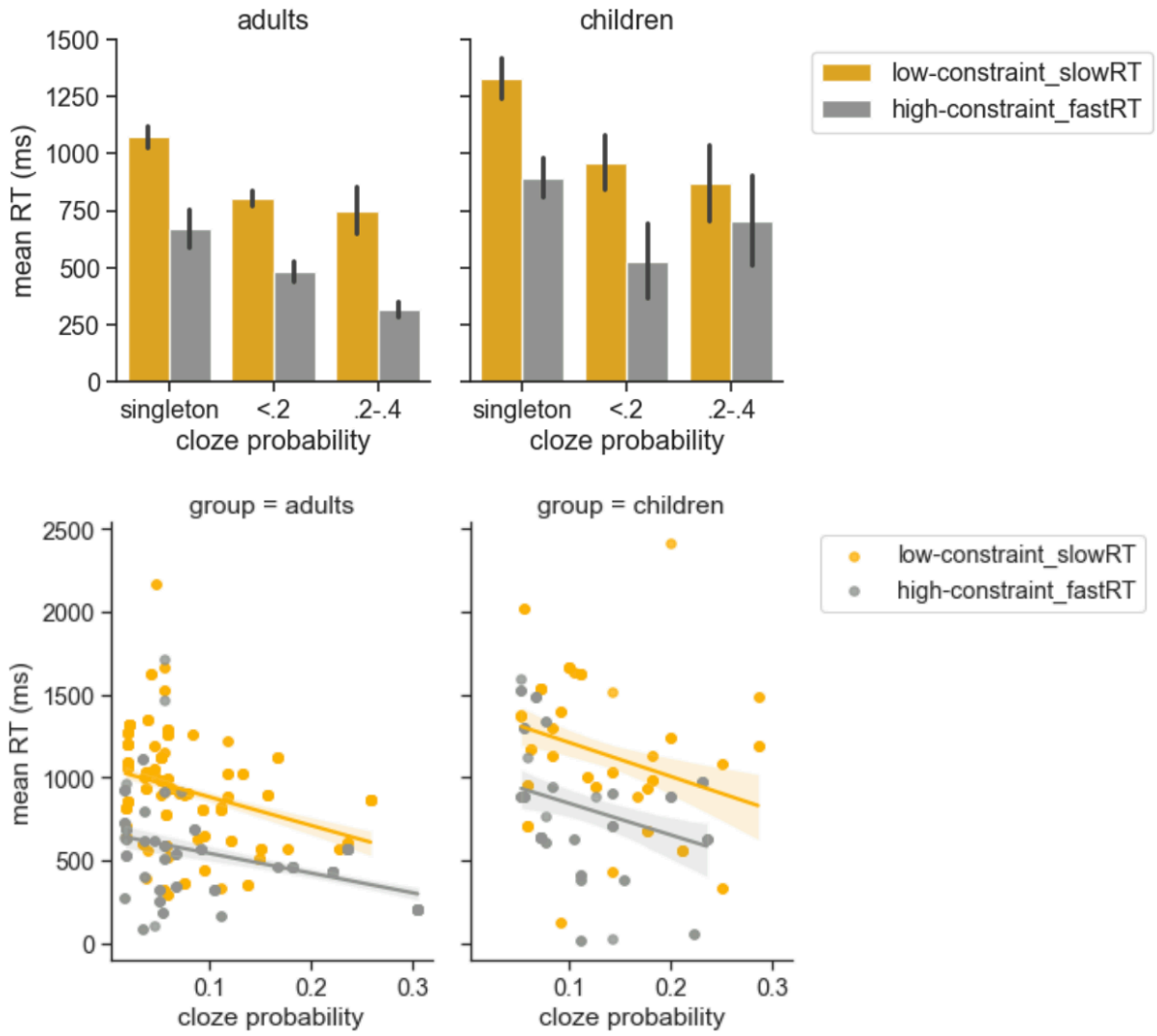


Figure 25: Relationship between sentence constraint and RT across different cloze probabilities. Error bars represent standard error of the mean.

A mixed effects model for each response RT, now including mean-centered log-transformed average RT of other responses as a fixed effect, instead of modal cloze probability,

revealed a significant main effect of cloze probability (β = -1.63, SE = .75, p = .03) and a significant main effect of item constraint (β = .98, SE = .11, p < .0001) with no interaction with group. Children and adults produced responses faster when the context had faster candidates overall, compared to when the average RT of responses was relatively slow.

The results show that having faster competitors yields faster responses at all levels of cloze probability, consistent with the race model's predictions, for both children and adults. In this respect, the speeded cloze responses reveal a shared predictive mechanism in children and adults which involves competition among multiple lexical candidates that race toward an activation threshold.

A remaining question is why the children did not show the same kind of relationship between sentence constraint and cloze RTs as found with adults when sentence constraint was measured by modal cloze probability. For adults, the two different ways of measuring sentence constraint yielded identical outcomes, whether we used modal cloze probability or mean RTs. Modal cloze probability characterizes a context in terms of the strength of fit between the context and the set of possible continuations, through the existence of a dominant candidate (however fast), while mean RT does so through the elicitation of fast candidates (however many there are). This indicates that for adults, a highly constraining context is typically a race with both a frequent winner and strong candidates, which leads to fast RTs overall. This was not true for children. The analysis with modal cloze probability revealed that children did not show a high-constraint and fast RT relationship at all levels of cloze probability, as the effect of sentence constraint on RT disappeared in the low-cloze responses.

We conducted race model simulations to model the kinds of contexts that could produce the children's speeded cloze response patterns which diverged from adults' (i.e., high modal cloze

probability but slow average RTs, or low modal cloze probability but fast average RTs). We show that the children's patterns can also be captured by the same basic assumptions of the race model but with different candidate profiles compared to adults, where having fewer or slower candidates leads to the unique RT patterns in children.

4.2.2.4 Race model simulations

In the simulations, a race context is defined through a set of candidates, where each candidate has a mean finishing time and a distribution, which together characterize the relative strength of each candidate given a particular context. In a single trial of a race, a single finishing time is sampled from each candidate's distribution, and the means are then ranked to determine the candidate with the fastest finishing time, i.e., the cloze response. Multiple trials are simulated for each race context. In this way, the win proportions of each candidate and its average winning time across the multiple trials can be calculated and compared across different race contexts.

We first replicated Staub et al.'s (2015) simulations, which demonstrated that a highly constraining context yields faster finishing times for candidates at all levels of cloze probability. The default race context was defined by 10 candidates with each of their mean finishing times, evenly distributed with 10 ms increments from 955 ms to 1045 ms (i.e., 955 ms, 965 ms, 975 ms, …, 1045 ms; $N = 10$) with a set variance of 50 ms. Then, highly constraining contexts were created by changing the default candidate set in two different ways: i) increasing the difference between the means for different candidates to 20 ms instead of 10 ms, or ii) adding a candidate with an exceptionally fast mean finishing time (i.e., 915 ms, 955 ms, 965 ms, 975 ms, …, 1045 ms; $N = 11$). A total of 100,000 trials were simulated for each race context. They showed that highly constraining contexts yield higher win proportions for the modal response and faster finishing times for all candidates in the race compared to the default context. The reason for this is that, in

both scenarios, the strongest competitor becomes harder to beat, and so the other candidates win the race only on their fastest trials.

In order to model races that could capture children's speeded cloze patterns, where high modal cloze probability did not always yield fast mean RT, we manipulated the competitors in Staub et al.'s highly constraining context in two different ways: i) by changing the competitors' mean finishing times, or ii) by changing the number of competitors. These manipulations were motivated by the observation of items where children notably produced fewer distinct responses than adults, as discussed in the previous section. First, in order to model the effect of competitor strength on race outcomes, we took Staub et al.'s highly constraining context with an exceptionally fast candidate and increased the speed of each candidate by 20 ms, except for the fastest candidate, which served as the target for comparison between the different contexts simulated. The resulting set of competitors in the new highly constraining context, therefore, had faster candidates (e.g., 915 ms, 935 ms, 945 ms, …, 1025 ms; N = 11) than in the original set. A race with slower competitors was created by adding 20 ms to each of the original candidate's mean finishing times (e.g., 915 ms, 975 ms, 985 ms, …, 1065 ms; N = 11). A total of 100,000 trials were simulated for two race contexts, and the results are plotted in Figure 26, along with Staub et al.'s default and highly constraining contexts.

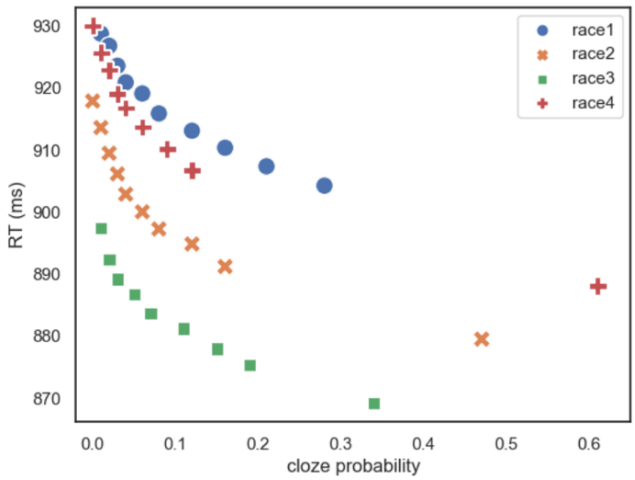


Figure 26: Results of race simulations modeling contexts with different speed of competitors

Note. Race 1: Staub et al.'s default context; Race 2: Staub et al.'s high constraint context; Race 3: Staub et al.'s high constraint context with faster competitors; Race 4: Staub et al.'s high constraint context with slower competitors.

Compared to the default context (Race 1), all of the highly constraining contexts (Race 2-4) yielded higher modal cloze probability, represented in the plot by the rightmost candidate in each race context. However, within the highly constraining contexts, there was variability in modal cloze probability depending on the speed of the competitors. Note that this modal candidate has exactly the same underlying finishing time distribution in each of these three simulations (Races 2-4). Compared to Staub et al.'s Race 2, the race representing a highly constraining context with faster competitors yielded faster finishing times overall but lower modal cloze probability (Race 3), while the race with slower competitors yielded slow finishing times overall but higher modal cloze probability (Race 4). The results demonstrate that the strength of competitors can lead to variation in modal cloze probability, which could explain children's speeded cloze patterns: a candidate can win more often despite being slow if its competitors are weak, hence producing a high modal cloze probability and slow mean RT pattern.

We also simulated race contexts that had half or double the number of competitors in Staub et al.'s highly constraining context with the exceptionally fast candidate. Again, 100,000 trials were simulated for each of the races, and the results are plotted in Figure 27.

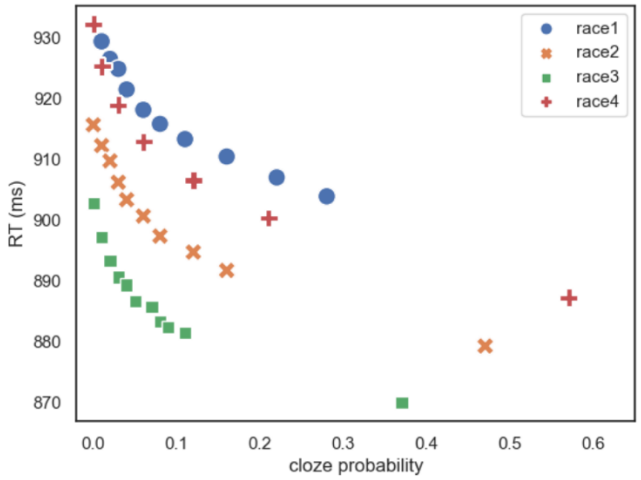


Figure 27: Results of race simulations modeling contexts with different number of competitors

Note. Race 1: Staub et al.'s default context; Race 2: Staub et al.'s high constraint context; Race 3: Staub et al.'s high constraint context with more competitors; Race 4: Staub et al.'s high constraint context with fewer competitors.

The results replicated the pattern observed in the previous simulation that manipulated the speed of the competitors. The race with more competitors elicited faster RTs but lower modal cloze probability (Race 3), while the race with fewer competitors produced slower RTs but higher modal cloze probability (Race 4).

The simulations combined show that the speed of competitors and the number of competitors can influence race outcomes, either of which yields the pattern observed in children's speeded cloze responses. Staub et al.'s simulations demonstrated that a highly constraining context could be a race where there is an exceptionally fast candidate, which yields faster overall winning times, due to a competitive race process. Our simulation results suggest that for children, the same

kind of competitive dynamics occur and that a highly constraining context for children can similarly be a race with a fast candidate but with weaker (slower or fewer) competitors in the race. This results in a pattern that diverges from adults: highly constraining contexts sometimes produce slow RTs, particularly in low-cloze responses that rarely win the race.

4.2.3 Discussion

The current study examined children's self-generated speeded cloze productions and whether they show the same hallmark patterns as adults, reflecting an underlying competitive dynamic, involving multiple lexical candidates racing toward an activation threshold to get produced (Staub et al., 2015). We designed a gamified version of the speeded cloze paradigm and collected responses from school-aged children and adults, taking advantage of visitor interest in a language-focused museum. The analysis of response rates and RTs revealed that while children's responses were slower than adults', they showed both key patterns that suggest the same competitive race mechanism in generating predictions in real-time. For adults, i) high-cloze responses were produced faster than low-cloze responses, at all levels of sentence constraint, and ii) highly constraining contexts produced faster responses than low constraining contexts, at all levels of cloze probability, supporting the race model. Children showed the same cloze and RT relationship but diverged from adults regarding the effect of sentence constraint on RT. However, we showed that this was due to an inappropriate measure of sentence constraint, and using an alternative measure resulted in identical patterns between children and adults, supporting parallel, competitive activation of multiple candidates based on context. Through race model simulations, we showed that the children's unique speeded cloze pattern could be a result of having slower or fewer competitors for a dominant candidate given a highly constraining context. We address these key findings in more detail below.

4.2.3.1 A race mechanism in children's real-time generation of next-word predictions

We expected two key patterns to appear in children's speeded cloze response RTs if children used a race-like process to generate candidates in a speeded cloze paradigm like adults. One was the relationship between cloze probability and RT, where high-cloze responses should be faster than low-cloze probability, regardless of how constraining the context is. This is because as candidates accumulate activation toward a threshold, faster candidates have a greater chance of reaching the threshold first and getting produced than slower candidates, resulting in faster responses for higher cloze probability. We found this relationship in children as well as adults, where high-cloze responses were overall faster than low-cloze responses, regardless of how constraining the context is. This pattern is consistent with a race-like process in generating predictions online which is shared between children and adults.

One might argue that the relationship between cloze probability and RT is also consistent with an alternative mechanism, where cloze responses are sampled from a probability distribution (Smith & Levy, 2011). The original sampling account, as discussed in the Introduction, does not make specific timing predictions, as it simply involves a one-time sampling process that was proposed to explain offline cloze probability. However, one could extend the theory to predict a meaningful relationship between cloze probability and RT, where it would have the same kind of relationship that is proposed to hold between cloze probability and subjective probability distributions. If people generate cloze responses by sampling from a distribution that reflects how good of a fit a candidate is given a context, that goodness of fit could be directly reflected in response times, in the same way as it is reflected in cloze probability. Stronger candidates would be associated with higher cloze probability and faster response times. In this way, the relationship between cloze probability and RT observed in both children and adults in the current study would

be compatible with a shared mechanism where cloze responses are generated by sampling from prior distributions.

While the relationship between cloze probability and response RT might be captured by the two alternative mechanisms, the second key finding, where highly constraining contexts yielded faster responses than less constraining contexts, supports the race model over a sampling account. According to the race model (Staub et al., 2015), the relationship between sentence constraint and RT is a consequence of a competitive race process where a candidate must be faster than its competitors to get produced, so that even non-dominant candidates must be faster to beat stronger competitors in a highly constraining context. The adults' speeded cloze responses in the present study showed the expected pattern: highly constraining contexts produced faster responses than less constraining contexts, and this effect did not interact with cloze probability, replicating Staub et al.'s (2015) findings and confirming the model's predictions. The children's responses diverged from those of adults, when sentence constraint was measured by modal cloze probability. However, using a more direct measure of sentence constraint, specifically the mean RTs of all other responses produced for a context, excluding the RT of the word itself, revealed the expected relationship between constraint and RT in both children and adults. Children's speeded cloze response patterns are, therefore, consistent with a race-like process in generating next-word predictions, where multiple candidates compete with each other to reach a threshold first and get produced. On the other hand, the constraint-RT relationship is difficult to explain with a sampling account, as it is not straightforward how the relative strength of competitors should affect the RT and cloze probability of a candidate given a context.

We additionally ran race model simulations in order to examine what kind of candidate profiles would yield children's speeded cloze patterns that diverged from adults, where contexts

with high modal cloze probability did not consistently yield fast mean RT and vice versa, low modal cloze probability did not always represent a context with slow responses. The simulation results showed that either reducing the number of competitors or increasing the time it takes for competitors to reach threshold reproduced the children's speeded cloze RT patterns. This suggests that children's highly constraining contexts may be contexts that have fewer or slower competitors for a dominant candidate compared to adults'. Importantly, the results suggest that the divergence between children and adults can be explained by a difference in candidate profiles rather than by a fundamentally different predictive mechanism. Children's response patterns naturally follow from a race mechanism proposed to explain adults' speeded cloze behavior.

Given the way the participants were recruited in a live museum setting, with voluntary participation, our study included a relatively broad range of school-aged child participants. As cloze probability and RTs are group-based measures, it is difficult to observe the effect of age by treating it as a continuous variable for the groups combined, as that would involve calculating cloze values by collapsing responses from adults and children together, which would potentially obscure any differences in response profiles between them. This makes it challenging to pinpoint whether the different candidate profiles observed between the children and adults in the current study emerge with increasing age or specifically through the development of vocabulary knowledge (e.g., Borovsky et al., 2012). One possibility is that the development of vocabulary and world knowledge leads to more dominant candidates or multiple strong candidates in highly constraining contexts, yielding the adults' response patterns where modal responses have higher cloze probability and responses are overall fast. Another possibility is that the speed of processing information given by the preceding context improves with age, which then quickly leads to dominant responses and faster response times given more informative contexts. In our additional

analyses, we included age as a continuous variable within the children's group, to examine whether it interacted with the key race patterns. We only found that cloze probability interacted with age, such that cloze probability had a stronger negative relation with RTs with increasing age. While the three-way interaction for cloze, constraint, and age within the children's group did not reach significance, this could be due to insufficient power, with limited total number of trials, or because the child-specific pattern we see is consistent across ages within school-aged children. While the current findings indicate that an adult-like race mechanism for generating parallel candidates during prediction is used by school-aged children, we encourage future work to examine the key factors that contribute to the change in children's candidate profiles to match adults'.

4.2.3.2 Theoretical contributions and implications

The main contributions of the current study to the existing literature on children's predictive processing are twofold. It offers an effective methodology to study predictions in children, and it helps to answer theoretical questions about the underlying cognitive mechanisms involved in children's predictive behavior.

The speeded cloze paradigm used in this study addresses some shortcomings of previous measures used to study children's predictive processing. Much of the literature has focused on whether or not children show predictive behavior like adults in comprehension measures like eye-movements in a visual world paradigm. Exhibiting adult-like patterns in these tasks may suggest that children can analyze the linguistic input and use it to correctly select or respond to an appropriate target given a set of presented options. However, since these experiment designs require the presentation of visual objects or linguistic input (in EEG/ERPs) which serve as the target or distractors to which responses are measured, it is difficult to directly probe whether children generate multiple candidates in parallel based on context alone. Our findings show that

the speeded cloze paradigm offers a more direct way of examining children's predictive processes, where response latencies offer informative evidence in support of a competitive race-like mechanism in generating predictions online. We are currently conducting a follow-up study using a reading-free version with auditory context presentation, which makes it possible to test younger children.

The results speak to the literature on prediction (Altmann & Mirković, 2009; Van Petten & Luka, 2012; Dell & Chang, 2014; Huettig, 2015; Kuperberg & Jaeger, 2016; Luke & Christianson, 2016; Pickering & Garrod, 2013; Pickering & Gambi, 2018; Federmeier, 2007, 2022) and support the context-driven parallel activation of multiple candidates in real-time prediction (Roland et al., 2012; Staub et al., 2015; Frisson et al., 2017; Ness & Meltzer-Asscher, 2021). Consistent with Staub et al.'s (2015) proposal, the adults' and children's speeded cloze response time patterns in the current study support viewing prediction as a process where multiple candidates accumulate activation in parallel based on preceding context. The relative strength of a candidate given a context is best reflected in cloze response latencies, which reflect how quickly it reached the activation threshold, relative to other candidates generated by the context. More constraining contexts lead to modal responses with higher cloze probability and faster response times, because those contexts generate faster candidates than less constraining contexts, potentially due to more abundant or useful information that can be used to generate stronger predictions.

The fact that children's speeded cloze responses in the current study could be captured through the same race process proposed to explain adults' predictive behavior suggests that the same mechanism is used from a young age rather than developed later in life. The results challenge alternative theories for explaining how children generate candidates, such as through a non-parallel activation of a single candidate or by sampling candidates from a probability distribution (e.g.,

Smith & Levy, 2011), which cannot capture the systematic relationship between sentence constraint and RT observed with children as well as adults. Although the child participants in the current study were older than the age group typically examined in previous studies, their specific speeded cloze patterns suggest that the race-like predictive mechanism is not unique to adults and is present in school-aged children, and an overall slowdown in RTs at younger ages, for example, is unlikely to affect the main patterns observed with the children and adults in the current study. Our findings can also provide new insights to theories that claim prediction is a driving factor of children's language development (i.e., error-based learning: Elman, 1990; Chang et al., 2006; Dell & Chang, 2014; Fazekas et al., 2020; Peter et al., 2015; see also Rabagliati et al., 2016; Lidz et al., 2017), and future work could investigate how a race mechanism in generating candidates during real-time prediction affects children's acquisition of new linguistic knowledge.

The current study additionally highlights the benefit of conducting research outside a lab setting where more active interactions with the participants and the general public was possible. The experiment setup at the language museum worked remarkably well. We were able to collect good spoken language data, the children and parents enjoyed participating in the gamified experiment, and the students who led the sessions on-site benefited from their training. The museum found that it enriched their visitors' experience. This opens useful avenues for testing large numbers of participants outside a lab setting. Importantly, it depended on very careful attention to designing the study and training the student researchers, so as to create a positive experience.

4.3 Conclusion

We used a speeded cloze paradigm to probe the underlying predictive mechanisms in school-aged children as well as adults. We examined whether children's speeded cloze responses showed adult-

like patterns that reflect a competitive race dynamic, where multiple candidates accumulate activation based on context and compete toward a threshold (Staub et al., 2015). We recruited child and adult visitors at a language-focused museum and administered a gamified version of the speeded cloze task, collecting self-generated responses and RTs. The analyses of the speeded cloze data augmented with computational race simulations showed that children's response patterns can also be captured by a race mechanism in generating candidates for upcoming words during real-time sentence processing, although for some highly constraining contexts, children may generate fewer competitors or weaker competitors for a dominant candidate compared to adults, leading to RT patterns that diverge from adults.

Chapter 5: The timing of generating expectations in adults and children

5.1 Introduction

A key factor that contributes to successful real-time sentence processing is the ability to rapidly construct and update representations moment-by-moment, using information that becomes available as a sentence unfolds. A large body of psycholinguistic literature has shown that comprehenders actively and rapidly build complex linguistic structures and compositional meanings and further use those representations to generate their own expectations for upcoming inputs (Altmann & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007, 2022; Hale, 2001; Levy, 2008; Van Petten & Luka, 2012; Hickok, 2012; Pickering & Garrod, 2013; Pickering & Gambi, 2018; Huettig, 2015; Kuperberg & Jaeger, 2016; Ryskin & Nieuwland, 2023). Importantly, studies have shown that children as young as two years old demonstrate this kind of anticipatory behavior and show similar predictive patterns empirically as adults. Despite the parallels in predictive processing observed between children and adults in previous work, it is less clear whether children exhibit adult-like capacity to accurately parse and use different types of available contextual information, in combination with their world knowledge, to rapidly constrain their expectations for incoming words during sentence processing.

In the present study, we focused on children's sensitivity to argument role information in sentence contexts when generating predictions for verbs in real-time. Argument roles in verb prediction serves as a good test case for examining the representations and expectations hat children, as well as adults, can rapidly build incrementally and what types of information they can effectively use in prediction, as it helps to disentangle the semantic and syntactic information that can be used to generate potential candidates for upcoming words. To take an example, between

sentences a) and b), the individual lexical items are identical in both sentences, and the only difference is the order in which the arguments are introduced.

a. The restaurant owner forgot which customer the waitress had *served* last night.

b. The restaurant owner forgot which waitress the customer had *served* last night.

The difference in word order, in English, yields distinct thematic role interpretations, where the two arguments *customer* and *waitress* take on different thematic roles as either the agent or patient. Critically, this difference between the two contexts makes the verb *served* plausible and highly expected in sentence a) but contextually inappropriate in sentence b), given the world knowledge about the types of events waitresses and customers typically engage in, each with their respective roles. The question is to what extent the verb *served* is expected given the role-reversed context in sentence b), where based on semantic associations of all of the preceding words, *served* is highly related, but given the argument structure, it is contextually inappropriate.

Previous work on adult's verb prediction using argument roles have shown mixed results depending on the measures used to probe predictions. Studies using the N400 ERP component as a measure of how much a target word was predicted have reported that people initially consider a contextually inappropriate but semantically associated role-reversed verb to the same extent as a role-appropriate verb, suggesting that argument role information does not limit fast predictions to contextually fully compatible candidates (Kuperberg et al., 2003; Hoeks et al., 2004; Kim & Osterhout, 2005; Chow et al., 2016, 2018). However, more recent studies using the speeded cloze paradigm have shown that people rarely produce role-reversed verbs as cloze completions even under time pressure and that role-reversals are produced with longer response latencies than role-appropriate verbs, indicating that they are not strongly predicted (Chow et al., 2015; Nakamura et al., 2024). While determining the exact cause of the diverging results from these different measures

is beyond the scope of the current study, here, we focused on testing children's ability to rapidly parse argument roles and constrain their verb expectations using argument role information, by examining their performances on various versions of speeded cloze tasks, which require quickly generating sentence continuations based on sentence contexts.

Previous studies on children's understanding and parsing of argument structure have shown that the knowledge of argument roles is developed from a young age. By around two years old, children can use word order to determine the agent and patient of an action described with a novel verb in reversible transitives (Gertner et al., 2006), and they can generalize thematic roles across different structures to correctly determine the agent of a novel verb, despite the difference in the sentence structure (Fernandes et al., 2006). Even younger 19-month-old infants have shown to use their knowledge of argument structure to infer the meaning of novel verbs, at least in cases where verb subcategorization requirements are satisfied (Lidz et al., 2017). Studies with older children have also reported instances of thematic role priming that occurs across different known verbs, even when other confounding variables like the animacy and discreteness of the arguments are controlled for (Thothathiri & Snedeker, 2011). These studies indicate that children's knowledge of argument structure is robust from an early point in development, such that it can be generalized across different contexts and utilized for processing new word meanings.

While the knowledge and real-time parsing of argument structure appear to be robust from a young age, it is less clear whether children actively use that information in real-time prediction. Gambi and colleagues' (2016) visual world eye-tracking study with 4-year-olds suggests that children use argument structure, which is indicated through word order in English, to generate hypotheses about the likely patient when given an agent and the event. In the study, children showed reduced looks toward the agent and increased looks toward the patient in advance of

hearing the complete sentence that contained the full argument role information, which occurred more when the agent and verb context was predictive of the patient (*Pingu will ride the…*) compared to when the context was not informative (*Pingu will pull the…*). While these results suggest that children actively use such contextual information in prediction, there are two caveats. One is that their findings with adults diverged from what Kukona et al. (2011) observed with adults using a similar paradigm. In Kukona et al.'s study, the adult participants initially looked at the agent and patient with similar proportions even when the agent role was already filled in the input sentence, suggesting that initially, predictions may not be restricted to only role-appropriate candidates but rather based on lexical associations. One limitation of the visual world paradigm, which was used in both Gambi et al.'s and Kukona et al.'s studies, is that the potential agents and patients as well as other distractors were presented on the screen, which can be considered and selected from. This makes it difficult to confirm whether looks toward an agent, for example, when the agent role is already filled by the preceding context, are driven by faulty predictions which were unconstrained by the argument roles provided in the sentence context, or whether there was some additional interference from the candidates that were explicitly presented on the screen. Re-examining the effects using an experimental paradigm that does not require explicitly presenting pre-selected candidates could, therefore, further clarify whether role-inappropriate candidates are generated based on sentence contexts.

Another factor to consider with respect to the prior studies on argument roles with children is that they have mostly examined cases where one can predict another argument given a verb and an argument (Gambi et al., 2016; Özge et al., 2019, 2022). It is unknown whether children show adult-like sensitivity to argument structure when the situation requires them to predict (un)likely verbs given a pair of arguments and their respective argument roles. Constraining verb

expectations to role-appropriate ones requires correctly assigning the agent and patient roles to the preceding arguments and then generating hypotheses for the kinds of events that involve those two participants with their respective roles. It is possible that this kind of process may be more difficult for children, especially as non-canonical structures like object relative clauses and passives are often considered a challenging structure to parse (Borer & Wexler, 1987; Bencini & Valian, 2008; Messenger, Branigan & McLean, 2011; Huang et al., 2013). Given the complexity of the structure, where a verb follows its two arguments, it is possible that children resort to lexical associations instead of using argument role information to limit candidates to only role-appropriate ones.

In the current study, we examined whether children and adults use argument role information provided by sentence contexts in order to generate contextually appropriate candidates for upcoming words. We focused on the extent to which frequently produced responses given particular argument role contexts were produced in contexts where the argument roles were reversed. We designed a child-friendly version of a typical speeded cloze task for Experiment 5-1 and created different versions of the same paradigm for Experiments 5-2 and 5-3, in order to address specific questions following each preceding experiment.

5.2 Experiment 5-1

5.2.1 Method

5.2.1.1 Participants

Participants were native English-speaking children (N = 67; 4-12 years old; mean age = 9) and adults (N = 157; 18-59 years old; mean age = 34). All participants voluntarily participated in the experiment as part of a visitor experience at a local language-focused museum.

5.2.1.2 Materials

The experiment stimuli consisted of 20 sentence fragments that were created using a templatic structure, *This is {NP1} that {NP2}*, and child-friendly vocabulary (Table 4). Each sentence fragment was presented in two conditions, canonical and reversed, where a high-cloze response in the canonical context (e.g., *This is the girl that the bee* ***stung***) would be an implausible, role-reversal response in the reversed context, in which the preceding arguments are swapped (e.g., *This is the bee that the girl* ***stung***.). In addition to the 20 critical items, there were 40 sentence fragments with varying constraints, which was determined by the cloze probability of the modal response and normed in a previous study (Block & Baldwin, 2010). These were included as critical items for another study which we do not discuss here. The combined set of items were divided into two presentation lists, each containing 20 argument role items (10 canonical, 10 reversed) and 20 high- and low-cloze filler sentences. Each participant saw one of the two lists which were randomized in order.

Table 4: Example experiment stimuli

Condition	**Sentence fragment**	**Example response**	**Estimated cloze probability**
canonical	This is the girl that the bee	stung	high-cloze
reversed	This is the bee that the girl	stung*	low-cloze

*Producing *stung* (i.e., the modal response in the canonical condition) in the reversed condition is considered a role-reversal error.

5.2.1.3 Procedure

A gamified version of the speeded cloze paradigm was created and administered using PCIbex (Zehr & Schwarz, 2018). Participants were told they would play a language science game called

'Race the Robot' and were instructed to read sentence fragments presented word-by-word and say aloud a continuation as quickly as possible. Instructions were provided in both verbal and written form, with three interleaved practice trials. Four practice trials preceded the experimental trials. Figure 28 presents the introductory slides with the instructions for the experiment leading up to the practice trials. All participants provided verbal consent and completed a demographic survey before beginning the experiment.

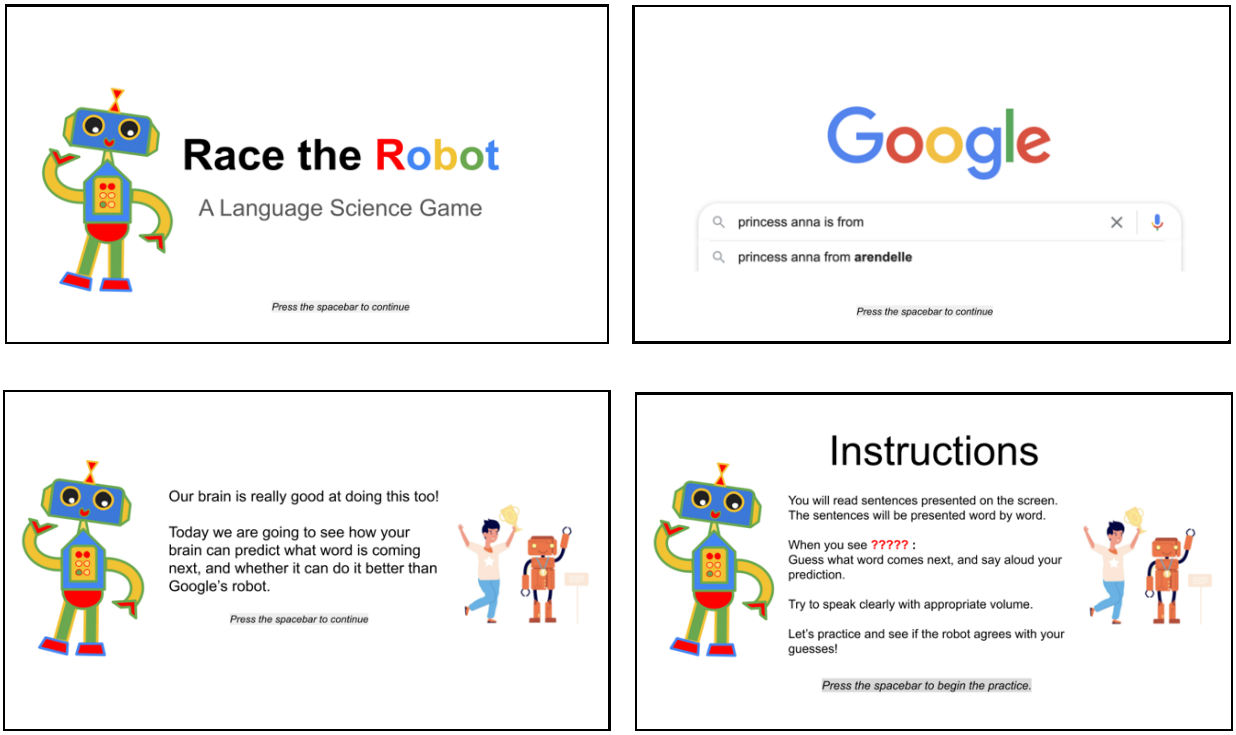


Figure 28: Illustration of the introduction slides in the gamified speeded cloze experiment 'Race the Robot' (presentation order: from left to right, top to bottom).

Each trial began with a '+' fixation mark presented for 800 ms, followed by a sentence fragment presented word-by-word with a 530 ms SOA (300 ms per word with a 230 ms blank in between). After the last final word in the sentence context, a series of question marks appeared in red font for three seconds, which prompted participants to produce a completion. The screen automatically moved onto the next trial after the three-second time limit. Figure 29 illustrates an

example trial. Participants' responses were recorded using a directional microphone robust to the surrounding noise in the museum. Each experiment session took about 20 minutes to complete.

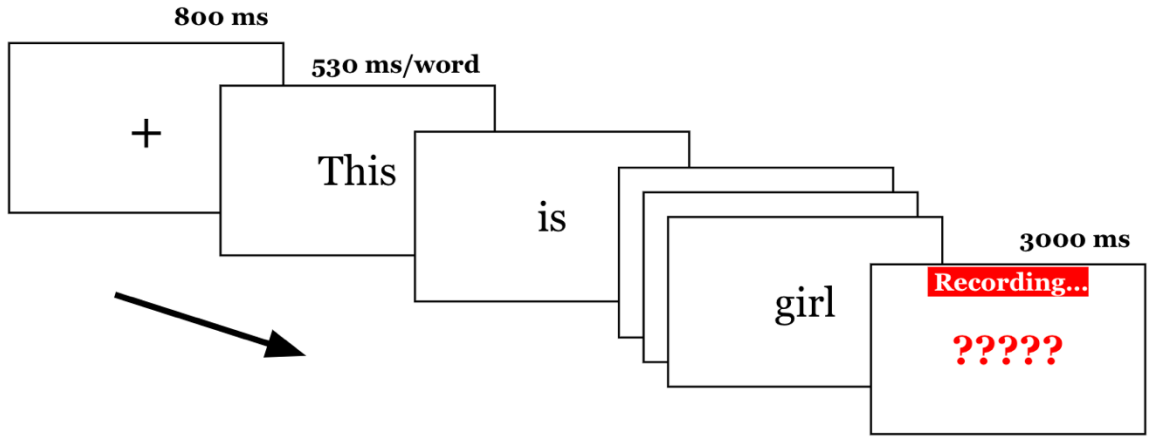


Figure 29: Illustration of a single trial in the speeded cloze experiment of Experiment 5-1.

5.2.1.4 Analysis

The collected audio files were pre-processed by obtaining automatic transcriptions using Google Cloud Speech-to-Text API and detecting automatic speech onset times using Chronset (Roux et al., 2017). The transcriptions and onset times were then manually checked and adjusted using Praat (Boersma, 2001). Responses that were unidentifiable or exceeded the 3-second time limit were removed from subsequent analyses.

Response rates and RTs were calculated with the pre-processed data. For each condition, we calculated the response rates and RTs of canonical modal responses. Canonical modal responses were the responses that were most frequently produced given the canonical context, for each item. The cloze percentage of those responses in the reversed condition represented role-reversal errors; all canonical modal responses were role-inappropriate in the reversed context, in which the argument roles were reversed.

Statistical tests for the main effects of interest were conducted through logistic and linear mixed-effects models using the lme4 package (v1.1.35.1; Bates et al., 2015) in R (v4.3.2; R Core

Team, 2023). A logistic mixed-effects model was constructed to examine group differences in role-reversal rates, by comparing response rates of canonical modal responses in the canonical and reversed conditions. The model included condition and group as fixed effects and initially included a maximal random effects structure for subjects, with the subject intercept and by-subject slopes for each of the fixed effects, and then simplified until the model reached convergence (Bates et al., 2013). Differences in RTs were examined with a similar structure using RTs with a linear mixed-effects model. The Box Cox procedure (Osborne, 2010) was applied to select a suitable transformation for analyzing RTs. All models included log-transformed RTs, unless otherwise indicated. A p value less than .05 for response rates and a t value greater than the absolute value of 2 for RTs were considered an indication of a significant effect.

5.2.2 Results

The results from Experiment 5-1 are presented in Figure 30, showing canonical modal response rates and mean RTs for both adults and children in the canonical and reversed conditions.

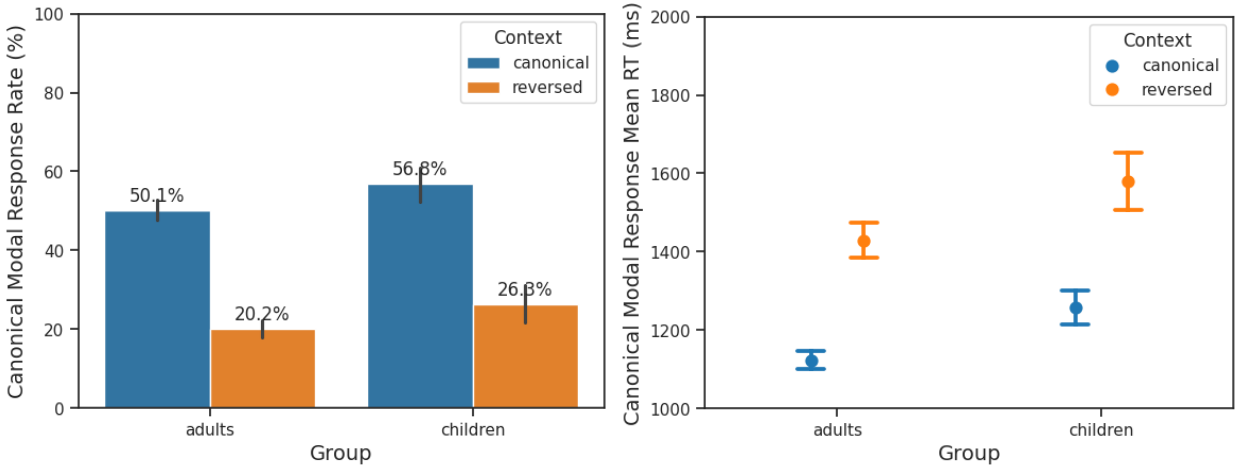


Figure 30: Canonical modal response rates (left) and mean RTs (right) of adults and children in Experiment 5-1.

The model for canonical modal response rates revealed a significant main effect of condition (β = -2.45, SE = .44, p < .001), indicating lower response rates in the reversed condition

compared to the canonical condition. There was no significant main effect of group (p = .44) or an interaction between condition and group (p = .08).

For canonical modal response RTs, results revealed significant main effects of condition (β = .27, SE = 0.05, t = 6.05), with longer RTs in the reversed condition compared to the canonical condition, and group (β = .24, SE = .06, t = 4.13), with children exhibiting longer RTs compared to adults. The interaction between condition and group was not significant (t = 1.34).

5.2.3 Discussion

The goal of Experiment 5-1 was to test whether children and adults exhibit similar sensitivity to argument roles when generating expectations for upcoming verbs. We designed a gamified version of the speeded cloze paradigm (Staub et al., 2015) in order to examine whether children and adults could rapidly use argument role information made available in the preceding context and produce role-appropriate sentence continuations under time pressure. The speeded cloze results showed that both children and adults avoided producing high-cloze responses that are role-appropriate in an argument role context when presented with the reversed context, where the argument roles were flipped. This pattern was found in both children and adults, with no significant differences between the two groups. Both children and adults also produced the canonical modal responses more quickly in the canonical context than in the reversed context, indicating that those responses were more readily available given role-appropriate contexts than role-reversed contexts. These results highlight children's and adults' sensitivity to both semantic and syntactic contextual information when generating expectations for upcoming words.

Before discussing the parallel between the two groups, we note that although our results indicate a significant difference in response rates and RTs between role-appropriate and role-reversed responses, the overall rate of role-reversal errors in the current study (about 20% for

adults) was higher than what is typically observed in speeded cloze studies (about 5%; e.g., Nakamura et al., 2024). We also found a higher average cloze percentage of the modal cloze responses in the canonical contexts than in previous studies (approximately 55% in the current study vs. 25% in previous studies). There were several differences between the previous studies and the current experiment, one being the fact that we used simpler sentence structures with easier vocabulary, in addition to carrying out the experiment on-site at a less controlled museum setting.

In order to examine whether the overall increase in role-reversal error rates was due to the experiment having been carried out in the museum, we conducted a follow-up experiment online, with adults recruited on Amazon Mechanical Turk. Participants were 180 adult native speakers of English (20-59 years old; mean age = 39). The same methods were used for stimuli presentation and data analyses. Figure 31 illustrates the response rates and mean RTs of canonical modal responses for the online adult participants.

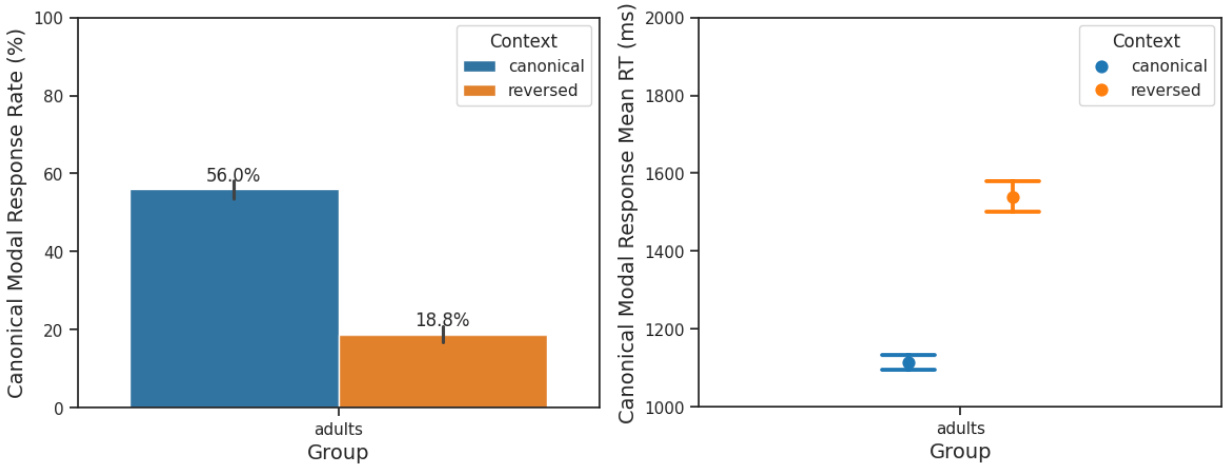


Figure 31: Canonical modal response rates (left) and RTs (right) of adults in an online replication of Experiment 5-1.

The model for response rates revealed a significant main effect of condition ($\beta = 2.56$, $SE = .20$, $p < .001$), with higher response rates in the canonical condition compared to the reversed condition. For (reciprocal transformed) RTs, a significant main effect of condition was observed

(β = -.31, *SE* = .05, *t* = -6.24), showing shorter RTs in the canonical condition compared to the reversed condition. The results overall replicated the patterns found with the museum participants, regarding the differences between canonical and reversed conditions as well as the average role-reversal error rate, which was about 20% for adults for both online and museum participants.

The results of the MTurk replication indicate that the increase in role-reversal errors in Experiment 5-1 was not due to the experiment having been carried out at a museum, as opposed to online or in an in-lab setting. Rather, it is possible that the increase was driven by the particular experimental sentences used in the current experiment, which were modified to fit the level of vocabulary and world knowledge of younger participants. The sentences were in templatic form (*This is NP1 that NP2…*), with limited additional context compared to the stimuli used in previous studies (e.g., Chow et al., 2018: *The restaurant owner forgot which customer the waitress had served last night.*). The sentence contexts in the current study also elicited sentence continuations immediately after the presentation of the second argument, as opposed to having an additional intervening word "had" in the previous studies. Given prior findings that suggest that the time available after the presentation of the arguments until the presentation of the verb modulates the role-sensitivity in EEG responses at the verb (Chow et al., 2018), it is possible that our stimuli provided even shorter time for argument roles to accurately constrain expectations compared to other stimuli that included an extra word before the presentation of the verb. It is also possible that the gamification of the speeded cloze paradigm, to fit the needs of the museum visitor experience affected the overall engagement in the experiment, which produced different results in the current experiment. While the exact reason behind the overall increase in role-reversal errors using the current set of materials and experimental design is beyond the scope of the current study, we

confirmed that the role-reversal rate does not significantly change depending on the experimental setting, whether the participants carry it out in a noisy museum setting or at home.

Returning to the discussion of the comparison between adults and children in the current experiment, a question was why we observed adult-like performance in children while previous studies have shown that children have difficulty in argument role processing. While it is possible that the age differences might matter, where our child participants were older than the children generally tested in prior work, another critical difference between previous studies and the current study was in the experimental paradigm used to test children's sensitivity. Specifically, psycholinguistic experiments with younger participants often include the presentation of pictures, and many experimental tasks involve selecting pictures that match the given linguistic inputs or producing sentences that accurately describe the given pictures. This is in contrast to the speeded cloze paradigm used in the current study, where participants read and produce continuations based on linguistic contexts alone, without additional visual contexts.

One reason to suspect that the contrast between text-only and the combination of pictures and audio might affect our ability to detect children's sensitivity to argument roles is related to the fact that children have shown to exhibit a weakness in revising initial commitments. The literature on children's language processing shows that children often fail to revise interpretations that they previously committed to (Trueswell et al., 1999; Huang et al., 2013; Bentea & Durrleman, 2018). In Trueswell et al. (1999), children were asked to listen and follow instructions which required them to move objects placed in front of them. When they heard a sentence like, "*Put the frog that's on the napkin in the box*," children were good at accurately parsing the relative clause and using the given information to distinguish between two frogs and carry out the correct action. However, children's performance significantly declined with sentences like, "*Put the frog on the napkin in*

the box," where children's initial interpretations were misled by the first prepositional phrase and failed to adapt to the new and correct interpretation that became available with the introduction of the second prepositional phrase. These findings suggest that while children actively form representations moment-by-moment based on preceding information, they have difficulty in revising those representations when they turn out to be felicitous with subsequent input.

If children indeed have difficulty in revising initial interpretations, it may be particularly challenging for children to generate contextually appropriate predictions when there are temporarily available interpretations that eventually turn out to be incorrect with subsequent information. This is especially relevant when examining children's sensitivity to argument roles using object relative clauses, where the first noun phrase that is introduced might be initially assigned an agent role which then must be switched to the patient role when the second noun phrase is introduced (e.g., *This is the bee that the girl…*). Previous studies have shown that children perform worse in interpreting passive sentences when they require revision of argument role assignments (Huang et al., 2013; Bentea & Durrleman, 2018). These studies have also tested children using pictures and audio, where the arguments are presented in front of the participants as they process the linguistic inputs. This is in contrast to the speeded cloze paradigm which requires participants to solely rely on the linguistic inputs to guide their interpretations and generate expectations for following inputs.

Based on the reviewed literature on children's sentence processing behavior, we hypothesized that the child participants were performing like adults in Experiment 5-1, because the task did not involve viewing pictures which could easily mislead children into felicitous interpretations that are difficult to recover from. Seeing a picture of a bee and a girl on the screen while hearing the sentence context, *This is the bee*…, could initially lock children's initial

interpretations as bee being the agent of the sentence. Under this scenario, successfully parsing the sentence would require recovering from an initial misparse and revising it when the bee turns out to be the patient, with the introduction of the second argument (*This is the bee that the girl…*). Having an image of a bee on the screen could make it more challenging for children to either prevent or rescue themselves from the initial agent interpretation of the first argument, compared to when only relying on the linguistic stimuli to build such representations.

In Experiment 5-2, we tested the possibility that the presence of images on the screen would worsen children's performance by creating another version of the speeded cloze task which involved presenting sentence contexts using both visual and auditory stimuli, similar to how children's linguistic knowledge is typically examined in psycholinguistic studies. We replaced text with pictures and audio for the presentation of sentence contexts and expected to see a contrast between adults and children, where children would show an increase in role-reversal errors compared to Experiment 5-1.

5.3 Experiment 5-2

5.3.1 Method

5.3.1.1 Participants

Participants were native English speakers, 114 adults (18-58 years old; mean age = 33) and 48 children (5-12 years old; mean age = 9), recruited from the same language museum as in Experiment 5-1. None of the participants had participants in Experiment 5-1.

5.3.1.2 Materials

The same experimental stimuli from Experiment 5-1 were used to create the visual and auditory stimuli for Experiment 5-2. The sentence fragments were recorded by a native English speaker, and pictures for each argument in the sentence contexts were selected from online sources.

5.3.1.3 Procedure

The overall experiment structure was identical to Experiment 5-1. Each trial began with a 800 ms fixation mark, followed by the presentation of two characters representing the arguments in the sentence context (1500 ms), and then a red box around the first argument that is mentioned, i.e., the patient, (1500 ms). Then, the audio recording of the sentence was presented and ended with a series of question marks indicating the 3-second time window, during which participants had to produce a cloze response. Figure 32 illustrates a single trial of Experiment 5-2.

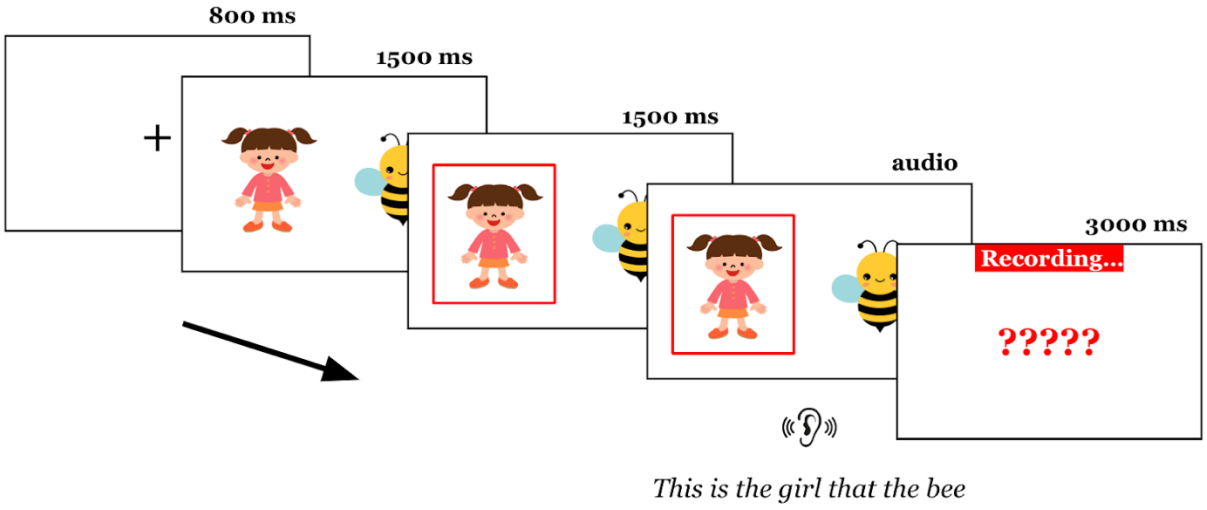


Figure 32: Illustration of a single trial in the gamified speeded cloze experiment of Experiment 5-2.

5.3.1.4 Analysis

The analysis was identical to Experiment 5-1.

5.3.2 Results

5.3.2.1 Experiment 5-2 results

The results from Experiment 5-2 are summarized in Figure 33, displaying the response rates and mean RTs for adults and children in canonical and reversed conditions.

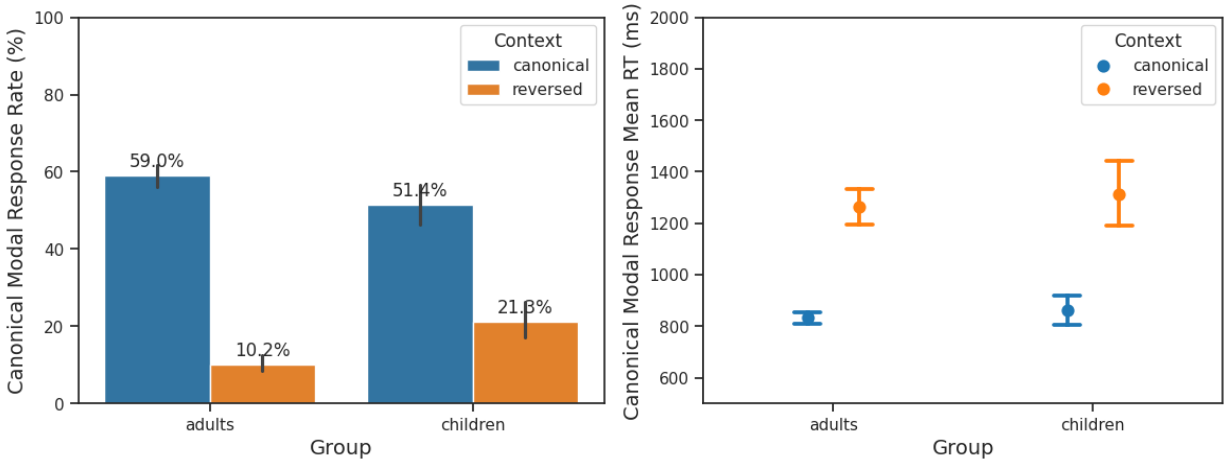


Figure 33: Canonical modal response rates (left) and RTs (right) of adults and children in Experiment 5-2.

The model for canonical modal response rates showed significant main effects of condition (β = -4.87, *SE* = .84, $p < .001$) and group (β = 1.16, *SE* = .47, $p = .01$), as well as a condition × group interaction (β = 2.72, *SE* = .75, $p < .001$). Post-hoc comparisons revealed that adults exhibited significantly lower response rates compared to children in the reversed condition ($p = .004$).

The model for canonical modal response RTs indicated a significant main effect of condition (β = .66, *SE* = .10, t = 6.45), with shorter RTs in the canonical condition compared to the reversed condition. There was no significant effect of group (t = -.90) nor a condition × group interaction (t = .35).

5.3.2.2 Comparison between Experiment 5-1 and 5-2 results

This section presents the results of the statistical comparisons between children's and adults' performances in Experiments 5-1 and 5-2.

A logistic mixed-effects model was used to examine the effects of condition (canonical, reversed), group (adults, children), and experiment (Experiment 5-1, Experiment 5-2) on canonical modal response rates. Results for canonical modal response rates revealed significant main effects of condition (β = -3.16, *SE* = .47, $p < .001$), group (β = .60, *SE* = .24, $p = .01$), and experiment (β = -.54, *SE* = .18, $p = .002$). Additionally, significant interaction effects were found for condition × group (β = 1.50, *SE* = .35, $p < .001$) and condition × experiment (β = -1.26, *SE* = .34, $p < .001$). The three-way interaction between condition, group, and experiment was not significant ($p = .11$).

A linear mixed-effects model was used to examine the effects of condition (canonical, reversed), group (adults, children), and experiment (Experiment 5-1, Experiment 5-2) on canonical modal RTs. Results revealed significant main effects of condition (β = .48, *SE* = .05, t = 9.89), with longer RTs in the reversed condition compared to the canonical condition, and experiment (β = -.42, *SE* = .06, t = -7.15), indicating shorter RTs in Experiment 5-2 compared to Experiment 5-1. The interaction between condition and experiment was also significant (β = .41, *SE* = 0.10, t = 4.30), suggesting that the difference in RTs between the canonical and reversed conditions varied across experiments. Additionally, a significant interaction between group and experiment was observed (β = -.37, *SE* = .12, t = -3.15), indicating that the difference in RTs between adults and children varied across experiments. No significant main effect of group (t = 1.40), or condition × group (t = .94) or condition × group × experiment (t = -.50) interactions were found.

We constructed separate models for children and adults, in order to further examine the differences between Experiments 5-1 and 5-2 within each group.

For adults' canonical modal response rates, results showed significant main effects of condition (β = -4.42, *SE* = .62, $p < .001$) and experiment (β = -.85, *SE* = .22, $p < .001$), with higher rates for the canonical condition than the reversed condition and for Experiment 5-2 than Experiment 5-1. Additionally, a significant interaction between condition and experiment was observed (β = -2.02, *SE* = .44, $p < .001$). Post-hoc analyses showed that canonical modal response rates were significantly lower in Experiment 5-2 compared to Experiment 5-1 in the reversed condition ($p < .001$).

Results for children's canonical modal response rates revealed a significant main effect of condition (β = -2.60, *SE* = .46, $p < .001$), showing lower response rates in the reversed condition compared to the canonical condition. Unlike with adults, children's response rates did not show a significant main effect of experiment (p = .29), nor an interaction between condition and experiment (p = .32).

For adults' canonical modal RTs, results showed a significant main effect of condition (β = .44, *SE* = .05, t = 8.01), with longer RTs in the reversed condition than the canonical condition. A significant main effect of experiment was also found (β = -.20, *SE* = .09, t = -2.25), indicating shorter RTs in Experiment 5-2 compared to Experiment 5-1. Additionally, a significant interaction between condition and experiment was observed (β = .44, *SE* = .11, t = 4.13), suggesting that the difference in RTs between the canonical and reversed conditions varied between experiments. Post-hoc analysis showed a significant difference between experiments in the canonical condition, with Experiment 5-2 showing shorter RTs compared to Experiment 5-1 (t = 5.18), but not in the reversed condition (t = -.18).

Results for children's canonical modal RTs revealed a significant main effect of condition (β = .56, *SE* = .10, t = 5.39), with longer RTs in the reversed condition compared to the canonical

condition. A significant main effect of experiment was also found (β = -.37, *SE* = .18, *t* = -2.04), indicating shorter RTs in Experiment 5-2 compared to Experiment 5-1. Additionally, a significant interaction between condition and experiment was observed (β = .49, *SE* = .21, *t* = 2.35), indicating that the difference in RTs between the canonical and reversed conditions varied between experiments. Post-hoc analysis showed that Experiment 5-1 had significantly longer RTs compared to Experiment 5-2 in the canonical condition (*t* = 3.15) but not in the reversed condition (*t* = .54), similar to adults' RT patterns.

5.3.3 Discussion

The goal of Experiment 5-2 was to test children's and adults' sensitivity to argument roles during verb prediction when sentence contexts were presented through audio and accompanied by images of the arguments. Participants saw pictures of the two arguments introduced in the sentence contexts, and following the recordings of sentence fragments, they had to quickly produce a response that would be a good continuation of the sentence. Both children and adults showed sensitivity to augment roles in their response rates and speed, similar to Experiment 5-1. However, we found a significant difference between two groups in role-reversal error rates, where children performed worse than adults in avoiding role-inappropriate continuations. This is in contrast to the adult-like performance observed in Experiment 5-1, in which the same experimental stimuli were presented in text only.

The results from Experiment 5-2 suggest that children have significantly more difficulty than adults in using argument roles to constrain their expectations for upcoming verbs, in this particular speeded cloze paradigm. While the adults produced role-reversal errors up to about 10%, the child participants produced nearly double the rate of adults. Given that the key difference between Experiment 5-1 and 2 was the mode of presentation of the sentence contexts, these results

suggest that child-adult contrasts arise depending on how the arguments and their roles are introduced.

While the results of Experiment 5-2 alone suggest that when sentence contexts are presented with pictures and audio, children and adults show different sensitivity to argument role information, it is difficult to conclude what caused the adult-child contrast which was not observed in Experiment 5-1 with text-only presentation mode. This is because our hypothesis that using pictures would worsen children's performance was not borne out. That is, the change in the experimental design led to an adult-child contrast because it improved the adults' performance rather than affecting the children's behavior. The between-experiments comparison within each group indicated that while children did not show different role-reversal error rates between the two experiments, the adults performed significantly better in Experiment 5-2 than in Experiment 5-1. The findings combined suggest that presenting sentence contexts with pictures and audio in Experiment 5-2 helped adults to more successfully use argument role information to constrain their expectations, in ways that did not benefit children.

One possible explanation for the contrast between children and adults in Experiment 5-2 only is that, rather than the presence of pictures, what contributes to successful role-based prediction is having some information about the arguments before the role information becomes available. Specifically, in Experiment 5-2, pictures of the arguments were presented on the screen for three seconds before the audio containing the sentence context was played. This additional time with the visual information about the upcoming arguments could have been actively used by adults in constraining their expectations even before the audio recording with the sentence context began to play. To put it more concretely, when participants see a picture of a girl and a bee, for example, this could provide an opportunity to generate expectations for events that are related to girl and

bee, as well as possible relationships between them. Participants could even go ahead and generate potential verb candidates that describe those events involving the arguments shown through the pictures. Importantly, these processes can occur before actually hearing the sentence context which contains argument role information. This could explain the results from Experiment 5-2, where the early availability of argument information may have helped adults to generate role-inappropriate predictions, while children did not use the information to initiate the generation process early on like adults.

In Experiment 5-3, we explored the possibility that adults use contextual information as early as they become available, even before sentence contexts are provided, by testing adults' role-sensitivity using the same experimental paradigm and presentation method used in Experiment 5-2 but removing the presentation delay preceding the audio containing the sentence contexts with the argument role information. The goal was to examine whether reducing the time between the presentation of the pictures of the arguments and the introduction of the arguments' role information given by the linguistic stimuli would reduce adults' role-sensitivity and decrease their performance from what was observed in Experiment 5-2 (about 10% role-reversal error rate) and move it closer to the higher error rate observed in Experiment 5-1 (about 20% role-reversal error rate), where with text-only presentation mode, argument role information was given simultaneously with the arguments through the word order.

5.4 Experiment 5-3

5.4.1 Method

5.4.1.1 Participants

Participants were 75 adult native speakers of English recruited on Prolific (18-58 years old; mean age = 37). None of the participants had participated in Experiments 5-1 or 5-2.

5.4.1.2 Materials

The stimuli were identical to Experiment 5-1 and 5-2, with the same images and audio recordings used in Experiment 5-2.

5.4.1.3 Procedure

The presentation of the stimuli was identical to Experiment 5-2, except that the 1500 ms delay between the presentation of the pictures and the presentation of the red box was removed. In this experiment, the red box appeared simultaneously with the pictures of the two arguments. Therefore, in contrast to Experiment 5-2, where participants saw the pictures for 3000 ms before hearing the audio, in Experiment 5-3, the time was reduced to 1500 ms, and the red box appeared simultaneously with the presentation of the arguments, such that the first image after the fixation cross was the arguments, with the first argument mentioned in the sentence context indicated with the red box. An example trial in Experiment 5-3 is illustrated in Figure 34.

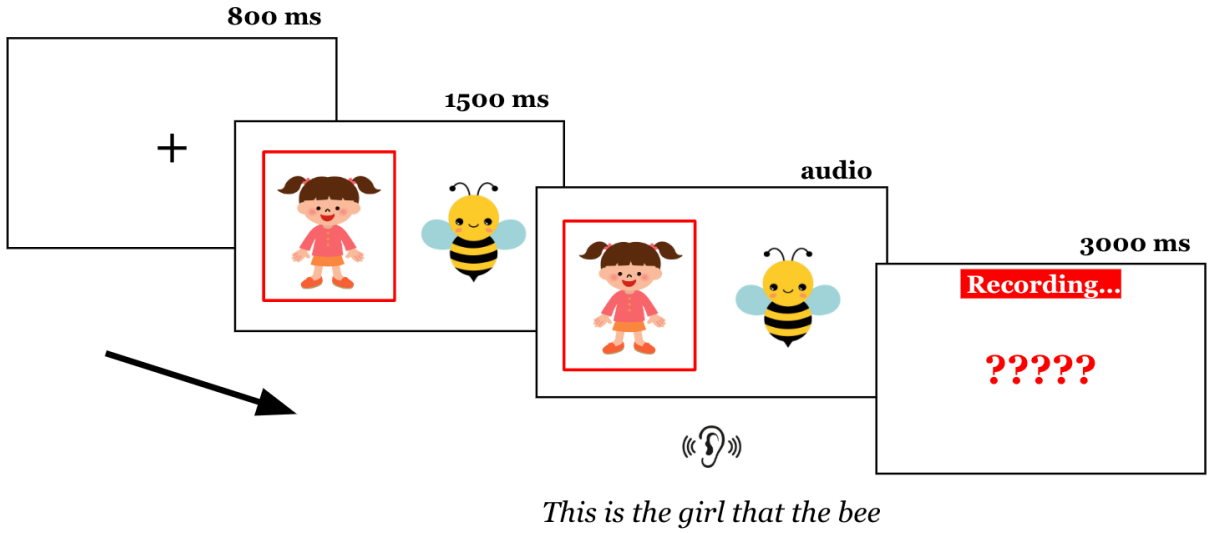


Figure 34: Illustration of a single trial in the gamified speeded cloze experiment of Experiment 5-3.

5.4.1.4 Analysis

The analysis was identical to the analysis for the MTurk replication of Experiments 5-1.

5.4.2 Results

5.4.2.1 Experiment 5-3 results

The results of Experiment 5-3 are presented in Figure 35, with response rates and mean RTs for adults in the canonical and reversed conditions.

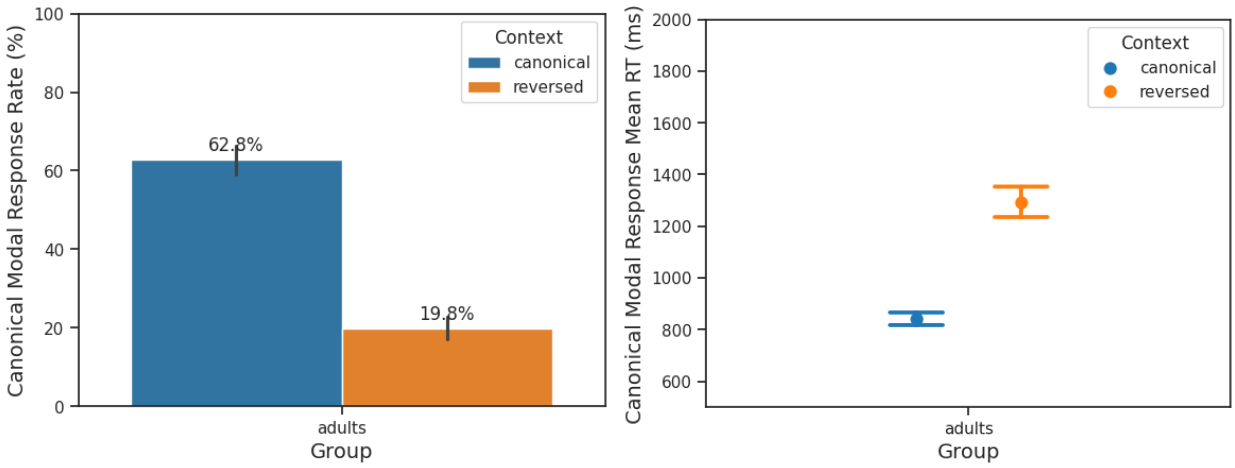


Figure 35: Canonical modal response rates (left) and RTs (right) of adults in Experiment 5-3.

For canonical modal response rates, a significant main effect of condition was observed (β = -2.54, SE = .26, $p < .001$), indicating higher response rates in the canonical condition compared to the reversed condition. For canonical modal RTs, a significant main effect of condition was found (β = .50, SE = .15, t = 3.48), with shorter RTs in the canonical condition compared to the reversed condition.

5.4.2.2 Comparison between Experiment 5-2 and 5-3 results (adults only)

We used separate models to compare the adults' performances between Experiments 5-2 and 5-3.

A logistic mixed-effects model was used to examine the effects of condition (canonical, reversed) and experiment (Experiment 5-2, Experiment 5-3) on canonical modal response rates. Results revealed a significant main effect of condition (β = -4.32, SE = .63, $p < .001$), indicating lower response rates in the reversed condition compared to the canonical condition. There was also

a significant main effect of experiment (β = .97, *SE* = .24, $p < .001$), with overall higher response rates in Experiment 5-3 compared to Experiment 5-2. Additionally, a significant interaction between condition and experiment was observed (β = 1.09, *SE* = .46, $p = .02$), suggesting that the difference in response rates between the canonical and reversed conditions varied between experiments. Post-hoc analyses indicated a significant difference in response rates between Experiments 5-2 and 5-3 in the reversed condition ($p = .002$), but not in the canonical condition ($p = .22$), where the adults' role-reversal error rate was higher in Experiment 5-3 than Experiment 5-2.

A linear mixed-effects model was used to examine the effects of condition (canonical, reversed) and experiment (Experiment 5-2, Experiment 5-3) on adults' canonical modal RTs. Results revealed a significant main effect of condition (β = .54, *SE* = .07, $t = 7.74$), with overall longer RTs in the reversed condition compared to the canonical condition. There was no main effect of experiment ($t = -.20$) or an interaction between condition and experiment ($t = -1.29$).

5.4.3 Discussion

The goal of Experiment 5-3 was to test adults' sensitivity to argument roles in generating candidates for upcoming words, when sentence contexts were presented auditorily along with pictures of the arguments. Participants quickly produced sentence continuations after seeing the pictures of two arguments and hearing sentence contexts which contained argument role information. The speeded cloze results indicated that the adults successfully avoided producing role-inappropriate responses most of the time and produced role-reversal errors at a rate of about 20%. This rate was similar to the rate observed in Experiment 5-1, where the same sentence contexts were presented as text, while it significantly diverged from the lower error rate found in Experiment 5-2, where the same sentence contexts were presented with audio and pictures but with

additional time between the pictures and the linguistic stimuli indicating argument role information. Below, we discuss the parallels and differences observed with adults across the three experiments.

The results of Experiments 5-1 and 5-3 revealed a similar rate of role-reversal errors for adults, which diverged from the adults' patterns in Experiment 5-2, where the adults demonstrated nearly half of the rate of role-reversal errors relative to the other two experiments. First, the comparable role-sensitivity found between Experiments 5-1 and 5-3 confirms that the mode of context presentation, i.e., whether it was through text or through audio and pictures, does not significantly affect the degree in which argument role information effectively constrains predictions. Processing the argument role information by reading sentence contexts word-by-word or by hearing the sentence while looking at images of the arguments did not modulate the adults' ability to use argument roles in their predictive processes.

The comparison of adults' performances between Experiments 5-1 and 5-3 combined and Experiment 5-2 revealed a meaningful contrast; the adult participants performed significantly better in Experiment 5-2, relative to the other two experiments. The key aspect of Experiment 5-2 which was different from the other two experiments was that information about the arguments were made available by the pictures which were presented for three seconds before the argument role information became available by the linguistic context. In Experiments 5-1 and 5-3, participants received the information about the arguments along with their thematic role assignments simultaneously, whereas in Experiment 5-2, participants saw the pictures of the arguments, without any thematic role information, before the sentence context that was presented with the audio recordings assigned the thematic roles to the arguments. The significant improvement of adults' performance in Experiment 5-2 indicates that having the availability of the role-independent information about the arguments, before their thematic role information,

significantly helped the adults in using the subsequently introduced role information to constrain expectations for upcoming verbs. We discuss the implications of this finding further in the General Discussion.

5.5 General Discussion

In three speeded cloze production experiments, we examined children's and adults' ability to generate predictions for upcoming words based on preceding contextual information, particularly focusing on the use of argument roles. We found across all of the experiments that both children and adults exhibited sensitivity to argument roles and avoided producing responses that were role-inappropriate the majority of the time. However, we also found that depending on how the information of the arguments and their thematic role information became available, children and adults showed distinct capacities in producing role-appropriate responses. In the sections below, we focus on the children's findings as well as the comparison across experiments and discuss implications for child-adult contrasts in argument role sensitivity and anticipatory processing.

We found clear evidence in Experiments 5-1 and 5-2 indicating that school-age children can rapidly process non-canonical structures, i.e., object relative clauses, assign correct thematic roles to incoming arguments, and generate contextually appropriate continuations based on information that is made available by the preceding context. Regardless of whether the sentence contexts were presented word-by-word as text (Experiment 5-1) or as audio recordings accompanied by pictures of the arguments mentioned in the linguistic stimuli (Experiment 5-2), children demonstrated the ability to correctly assign the thematic roles to the introduced arguments and use it for prediction. The child participants produced role-reversal errors about 23% of the time in the two experiments, which was not meaningfully higher than the adults' error rates found in Experiments 5-1 and 5-3. When children did produce role-reversal errors, those responses were

slower than when they produced those same responses in the correct, role-appropriate contexts. These findings combined suggest that school-age children can rapidly process argument role information and use it to generate role-appropriate candidates for the upcoming verb.

Our findings with children augment previous works that report sensitivity to argument structure in children's predictive processing (Gambi et al., 2016; Özge et al., 2019) and provide novel implications, particularly with regard to two aspects. First, while the earlier studies on children's processing of argument roles examined the ability to predict appropriate thematic role assignments for upcoming arguments using the contextual information indicating the relationship between the other argument and the verb (Gambi et al., 2016) or case-marking on the first argument (Özge et al., 2019), our findings further show that children can actively deploy their knowledge of argument structure in order to generate predictions for the verb, based on information about a pair of arguments and their respective roles assigned by word order (in English). The results indicate that school-age children can effectively use this type of information in predictive processing, even when it involves correctly parsing more complex, non-canonical sentence structures, like object relative clauses (Borer & Wexler, 1987; Bencini & Valian, 2008; Messenger, Branigan & McLean, 2011).

In addition, our findings contribute to the scope of methodology that can be used to examine predictive processes in children. Previous studies that have investigated anticipatory processing with younger participants as well as adults have typically used comprehension paradigms which require the explicit presentation of candidate continuations. For example, the visual world paradigm, which is widely used to study predictive processing in children, requires presenting pictures of multiple potential target objects that the sentence describes. While looks to the target picture before it is explicitly mentioned in the linguistic input can clearly indicate

predictive looks, it is difficult to determine to what extent the earlier looks were driven by the information provided by the linguistic context or the presentation of the pictures before the sentence contexts were presented. We took advantage of the speeded cloze paradigm, which does not require presenting candidate continuations and allows us to examine the candidates that children generate solely based on the given contexts, providing a more direct measure of self-generated expectations. The results using different versions of this paradigm showed that children, like adults, are most of the time able to use argument role information to guide their expectations for upcoming words.

Moreover, while we had initially hypothesized that the presentation of arguments as pictures would reduce children's performances as it could exacerbate initial misinterpretations and make it difficult to recover from role-reversed interpretations, we did not find any significant differences in children's performances with or without the presence of pictures. This indicates that seeing the arguments through pictures does not make it more difficult for children to avoid role-reversal responses, i.e., responses that would be highly appropriate and expected given the opposite role contexts. The only meaningful difference between Experiments 5-1 and 5-2 was that both children and adults produced predictable responses slightly faster, in the canonical condition, when the contexts were presented in audio and with pictures than when they were presented only in text. It is possible that this represents a fundamental difference arising from the mode of presentation, but since we did not specifically match the length of the presentation of the sentence contexts between the two modes of presentation, we cannot rule out the possibility that the difference is simply driven by measuring responses at slightly different time points, i.e., starting the recorder later in one version than the other. We leave this question for future research.

Despite the role-sensitivity found in children's as well as in adults' response patterns, we found evidence indicating that both groups sometimes produce role-reversals, i.e., verb candidates that fit the opposite argument roles. While this corroborates the EEG findings demonstrating that adults' predictions are sometimes initially role-insensitive (e.g., Kim & Osterhout, 2005; Kuperberg et al., 2003, 2007; Hoeks et al., 2004; Kolk et al., 2003, van Herten et al., 2005; Brouwer et al., 2012; Chow et al., 2016, 2018; Nakamura et al., 2024), the near 20% error rate is more surprising when considering the 5-10% role-reversal rate typically observed in speeded cloze production studies (e.g., Chow et al., 2015; Nakamura et al., 2024). As mentioned in the Discussion section for Experiment 5-1, the replication with online participants indicates that the overall higher error rate is unlikely due to the fact that the experiments were carried out in a museum setting, rather than online or in a lab. If that were the case, we would expect to see reduced role-reversal errors in the online replications, which was not observed. Instead, it is possible that the simplified context and templatic structure of our experimental items, which stand in contrast to the more elaborate sentence contexts used in other studies, reduced the strength of argument roles as contextual cues for constraining expectations and led to divergent role-sensitivity. We leave this question to be explored in future work, as it goes beyond the scope of the current study.

Finally, an additional novel finding observed in the adults' response patterns was that depending on the design of the speeded cloze paradigm, adults exhibited different degrees of susceptibility to argument role-reversals. Across the three experiments, the adults' role-reversal error rates significantly reduced in Experiment 5-2, where the key difference between Experiment 5-2 and the other two experiments was that the arguments were introduced through pictures, prior to when the arguments could be assigned thematic roles based on the linguistic context that followed the presentation of the pictures. More specifically, in Experiment 5-2, participants had

1.5 seconds to simply look at the pictures of the arguments, before the arguments were differentiated by a red box appearing around the patient of the action, for another 1.5 seconds. This was followed by the linguistic context. In the other two experiments, participants received the information about the arguments and their roles nearly at the same time, whether through word-by-word presentation (Experiment 5-1) or through a red box appearing with pictures of the arguments before the explicit linguistic input (Experiment 5-3). The fact that the adults' role-reversal error rates reduced nearly double the rate observed in the other experiments suggests that the adults were able to use the available information about the arguments that became available through the pictures, in advance of the linguistic input, to begin narrowing down the pool of candidates that could serve as a continuation. This represents an active strategy in generating predictions at the earliest moment possible. Unlike the adults, the children showed a consistent vulnerability to role-reversals even with the additional time provided with the pictures of the arguments. The contrast between the adults and children (in Experiment 5-2) indicates that children do not use the information about the arguments even when they are available earlier through visual cues. This potentially suggests that children do not take as a proactive approach as adults in using available information to quickly start generating predictions.

Taken together, the speeded cloze results indicate that school-age children can rapidly parse and use argument role information embedded in non-canonical structures to form expectations about subsequent words. In most cases, they exhibit adult-like sensitivity to argument role information provided in sentence contexts, when generating predictions for upcoming verbs. However, children diverge from adults in using all available information that is available to initiate generation processes at the earliest point possible. While adults show significantly reduced role-reversal errors when (role-independent) information about the arguments is available through

visual contexts, children do not actively use the available information to start constraining expectations to contextually appropriate candidates.

5.6 Conclusion

In the present study, we examined children's and adults' ability to use argument role information in generating predictions for upcoming verbs. Three different versions of the speeded cloze paradigm were used to examine the rate and speed of participants' responses when they were given sentence contexts which were presented through text or through audio and pictures. The results combined demonstrate that children show adult-like sensitivity to argument roles when predicting upcoming words in a speeded cloze task. However, unlike adults, children do not benefit from the separated introduction of the arguments and the arguments' thematic role information, indicating that children's predictive mechanism, compared to adults', takes a more conservative approach in using available contextual information to constrain expectations at the earliest possible moment.

Chapter 6: Prediction mechanisms in human and artificial intelligence

6.1 Introduction

Humans rapidly make predictions when comprehending language. However, certain types of contextual information do not immediately impact predictions, and a well-studied case of this in the sentence processing literature involves argument roles.

Argument roles refer to the roles of participants that take part in the event described by a sentence, i.e., who is the agent (do-er of the action) and who is the patient (undergo-er of the action). Extracting this information from the sentence and using it with prior knowledge to predict which event is being described is a hallmark of real-time language understanding. For example, in (1a), the verb served is a highly expected continuation given the preceding context, whereas swapping the argument roles, as in (1b), makes the same verb no longer appropriate.

(1) (a) The customer that the waitress **served**

(b) The waitress that the customer **served**

Surprisingly, studies with human participants have shown that the roles assigned to the arguments by the structure do not immediately impact verb prediction, in contrast to the context-independent lexical meanings of arguments. Human comprehenders show similar initial responses to a verb when it appears in role-appropriate and role-reversed contexts (e.g., 1a vs. 1b) (Chow et al., 2016; Kim & Osterhout, 2005). This has been taken to indicate that argument roles have a delayed impact on verb prediction in human sentence processing.

Recent work has used paradigms from experimental psycholinguistics to evaluate language models' representation of syntactic and semantic knowledge, and language models trained on next-

word prediction alone have shown strong levels of correspondence with human behavioral and neural data. However, the extent to which they accurately encode and utilize structural information, such as argument roles, in relation to structure-independent word meanings, to determine sentence plausibility remains an open question. Previous work has explored whether models can distinguish between plausible and implausible sentences involving argument role manipulations (Ettinger, 2020; Papadimitriou et al., 2022; Wilson et al., 2023; Kauf et al., 2023). However, much of this research has focused on comparing full sentences rather than isolating the relationship between argument roles and the verb, often introducing confounding factors such as animacy. This makes it challenging to accurately assess models' sensitivity to argument role information.

In this paper, we take a new approach in evaluating role-sensitivity in large language models, by focusing on models' representations of verbs that appear in either plausible or implausible sentence contexts, where plausibility is determined based on the verb's compatibility with the preceding argument-role bindings. This approach draws insights from experimental work testing humans' role-sensitivity and therefore offers a more direct evaluation of language models' sensitivity to structural information in comparison to humans than previous studies. Additionally, testing language models that are trained on next-word prediction provides a fertile testing ground for determining whether the systematic predictive patterns observed in human empirical behavior naturally arise from statistical co-occurrences and a prediction objective, as opposed to additional human cognitive processes. In this way, directly comparing predictive processing between humans and models can help us better understand the mechanisms that underlie human language processing.

We adapt materials used in psycholinguistic studies evaluating humans' sensitivity to argument roles, which allows us to use carefully constructed minimal pairs of sentences which only differ with respect to argument roles, while controlling for other factors like animacy. This

serves as a rigorous test in examining models' ability to extract argument-role bindings based on sentence structure, as it requires models to go beyond simply learning relations between various arguments and verbs, i.e., between real-world events and participants that are likely to be involved in those events.

We compare model performance on two different types of argument role manipulations, in addition to a baseline condition which has shown to elicit immediate sensitivity in humans, as a way to more systematically compare human and model behavior.

Through three experiments, we find that i) language models show weak sensitivity to argument role information relative to role-independent argument meanings, similar to human initial prediction behavior, ii) models do not show the same consistency across different types of argument role manipulations as humans do, indicating a difference in the way argument roles are processed in models and humans, and iii) models' weak performance may not necessarily arise from inaccurate processing of argument roles. These results overall indicate that even if models are able to distinguish plausible and implausible verbs based on argument roles, to varying degrees of success, the lack of generalization across sentences that share the same structural relation suggests that the models do not use the same mechanism as humans to compute argument-verb relations.

6.2 Related Work

To evaluate language models' representations of argument roles, reversing the order of the verb's arguments is a common design, paralleling the stimuli in human experiments. Researchers then compare differences in the reversed and felicitous conditions, using various metrics from the models. There are two major issues with existing work that we address. First, existing work often

relies on the animacy of the verbs' arguments. Second, work using different metrics often offer conflicting conclusions.

Papadimitriou et al. (2022) claim language models are able to effectively make use of word order-related information when arguments are switched for verbs with transitive subjects and objects, reflecting these distinctions imposed by selectional constraints on the verb in their representations. For instance, the models they evaluated would represent *The chef chopped the onion* differently from *The onion chopped the chef*. For this evaluation, they automatically switch the order of arguments in naturalistic corpora. Thus, it is unclear if these positive results are based on properties of the lexical items (i.e. frequency, animacy) that are learned more easily from distributional information, or more abstract representations of argument roles.

A more reliable way to measure the linguistic capacity of language models is to effectively treat them as psycholinguistic subjects (e.g., Ettinger, 2020; Futrell et al., 2019) across a range of configurations (see reviews by Linzen, 2021; Mahowald et al., 2024; Pavlick, 2022). Work in this vein presents models with minimal pairs of sentences and analyzes differences in language models' responses to each sentence. Language models' sensitivity to a variety of phenomena been evaluated with this paradigm (e.g., Linzen et al., 2016; Warstadt et al., 2020; Wilcox et al., 2023). For argument roles specifically, Kauf et al. (2023) find they are able to distinguish plausible events from implausible ones, assigning higher probabilities to sentences like *The teacher bought the laptop.* as opposed to *The laptop bought the teacher.*, but only when one participant is animate and the other is inanimate. Given the ability of language models to handle animacy even in atypical settings (Hanna et al., 2023), it is possible that the results of both Kauf et al. (2023) and Papadimitriou et al. (2022) may be tapping into this ability rather than a generalized representation of argument roles.

Ettinger (2020) presented a suite of psycholinguistically motivated diagnostics for BERT; one of these tests was on *argument role reversals*, which was similar in spirit to some of Kauf et al. (2023)'s stimuli but only tested animate participants. This study had different conclusions, finding that BERT was indeed sensitive to these role-related contrasts, generating role reversals in appropriate contexts, but not on par with humans. Working with this dataset, Li et al. (2021) evaluate the probabilities the models assign to the sentence at individual layers and finds that they are not sensitive to the role reversal sentences. These studies all use different methods of evaluation. Ettinger (2020) queried sentence completions made by BERT, while Kauf et al. (2023) determined whether the language models assigned lower probabilities to the implausible sentence of the pair.

We take a different approach to examine language models' sensitivity to argument roles by replicating psycholinguistic experiments with multiple conditions designed to isolate humans' representations of argument roles. These experiments track human processing in real time and specifically examine participants' responses to verbs, which reflect how the representation of the sentence is built up. To tighten the link to whether models are making human-like judgments, we also examine the models' responses to the verbs rather than sentence-level metrics through behavioral and representational methods in Experiments 6-1 and 6-2.

Furthermore, one reason why Transformers are hypothesized to capture many empirical patterns in human sentence processing is that their attention mechanisms are able to efficiently keep track of long-distance dependencies (Ryu & Lewis, 2021). Despite findings localizing handling certain syntactic dependencies to individual attention heads (Clark et al., 2019; Jian & Reddy, 2023; Vig & Belinkov, 2019), little work has been done on connecting these measures to psycholinguistic findings. Ryu and Lewis (2021) specifically found an attention head that handled subject-verb agreement in GPT-2, which corresponded with human processing of these

dependencies. This approach has not been tried for argument roles in a more generalized setting. We do so in Experiment 6-3.

6.3 Psycholinguistic Data

We use materials from previous psycholinguistic experiments which were carefully constructed to evaluate human comprehenders' sensitivity to argument roles in real-time sentence processing. These stimuli sets were designed to compare electrophysiological responses to verbs that appeared in different sentence contexts, and the different conditions have shown to elicit distinct N400 amplitudes, a neural response taken to reflect how strongly a target word was predicted based on the previous context (Kutas & Hillyard, 1980).

We use the materials from Chow et al. (2016) and Kim and Osterhout (2005), and label the conditions as swap-arguments, change-verb, and replace-argument (Table 5). Both studies were conducted in English on native speakers.

Table 5: Example sentences (1 pair = 1 item) in each condition. The swap-arguments and change-verb conditions involve argument role manipulations, while replace-argument serve as a control. Humans show greater sensitivity in the replace-argument than in the swap-arguments and change-verb conditions.

Condition	Items	Plausible	Implausible
swap-arguments	120	The restaurant owner forgot which *customer* the *waitress* **served** during dinner yesterday.	The restaurant owner forgot which *waitress* the *customer* **served** during dinner yesterday.
change-verb	96	The hearty meal was **devoured** with gusto.	The hearty meal was **devouring** by the kids.
replace-argument	120	The secretary confirmed which *illustrator* the author had **hired** for the new book.	The secretary confirmed which *readers* the author had **hired** for the new book.

Both the swap-arguments and change-verb conditions include manipulations of argument roles and verb plausibility. In the swap-arguments condition, the two arguments preceding the verb in the plausible sentence are swapped to create the implausible sentence. In the change-verb condition, the verb form is changed to create the plausible and implausible sentences. Although the two conditions involve different changes, both have the same consequence: verb plausibility changes because of the way the argument(s) are assigned different roles, while the argument(s) that appear in the context remain the same (e.g., waitress-customer or meal).

In addition to the two role-related conditions, we also include a replace-argument condition (Chow et al., 2016), which involves replacing one of the arguments with an entirely different noun. This results in changing the argument meaning rather than argument roles, and this has shown to yield immediate predictability effects in human verb predictions, as opposed to the previous two conditions which both fail to elicit rapid sensitivity.

The key human empirical pattern to which we compare language models' is: weaker sensitivity to argument roles (swap-arguments & change-verb) compared to argument meanings (replace-argument).

6.4 Models & Experiments

We use the following pre-trained language models for our analyses: GPT-2 (small, medium, and large) (Radford et al., 2019), BERT (base-uncased, large-uncased) (Devlin et al., 2019), and RoBERTa (base, large) (Liu et al., 2019). Details of the model properties are included in Appendix 6.A.

These models were selected based on prior work comparing human language processing patterns with measures derived from language models. Recent studies have shown that smaller versions of GPT-2 fit human reading times better than larger models (Kuribayashi et al., 2023; Oh

& Schuler, 2023). Steuer et al. (2023) confirm these results, showing that larger Transformer language models perform better on syntactic and semantic generalization tasks than they do at predicting reading times relative to smaller models. We selected different model sizes in order to examine how scaling up or down affects comparability with human performance. Additionally, GPT-2 models are unidirectional while BERT models are bidirectional, but they have a similar number of parameters. By manipulating the context available to a comprehender while controlling for model size, we can more effectively compare proxies of real-time incremental processing from the GPT-2 models compared to offline measures with the BERT-style models.

All models were accessed through the transformers (Wolf et al., 2020) or minicons library (Misra et al., 2022), built to work with the Hugging Face API. Code and data are available at https://github.com/umd-psycholing/RoleReversalLM.

We carry out three experiments, evaluating language models' ability to differentiate plausible and implausible verbs given the sentence. We specifically focus on addressing the following questions: (i) Do the models show a human-like pattern across the different conditions? (ii) Are these contrasts reflected in the models' representations across the intermediate layers? (iii) Do patterns in the models' attention weights reflect argument role sensitivity?

6.5 Experiment 6-1: Surprisal Effects

One of the most well-established measures linking language models to cognitive hypotheses is surprisal, or the negative log probability of a word given context. Surprisal theory (Hale, 2001; Levy, 2008) states that the difficulty associated with processing linguistic information can be operationalized with this measure. Language model surprisal has shown to strongly correlate with both human reading times (Shain, 2024; Smith & Levy, 2013) as well as the N400 EEG response (Frank et al., 2013; Michaelov et al., 2024). Current Transformer models perform more effectively

than other methods of language modeling (Merkx & Frank, 2021), and this relationship with reading times has been established cross-linguistically (Wilcox et al., 2023).

6.5.1 Methods

For each item, we compute the **surprisal effect** at the verb. As human sensitivity to argument roles is often measured at the target verb, this allows us to make a direct comparison between humans and model-based measures of prediction.

Even if we might expect models to assign lower probability, and thus higher surprisal, to implausible continuations, it is important to determine the surprisal effect on individual items, following work on the targeted syntactic evaluation of language models (Marvin & Linzen, 2018; Wilcox et al., 2023). This allows us to quantify not just whether the model is successfully capturing distinctions between sentences, but to what extent it is able to do so. We operationalize this effect in Equation 1, such that $context_i$ and $context_p$ are implausible and plausible versions of the same context, respectively, and S_{LM} is the language model's surprisal in Equation 2.

1. $S_{LM}(verb, context_i) - S_{LM}(verb, context_p)$
2. $S_{LM}(w, c) = -\log_2 P_{LM}(w|c)$

Verb surprisal estimates were obtained with Equation 2, and the surprisal effect for each item was obtained by subtracting the surprisal of the verb in the implausible context from the plausible context in all experimental conditions. Therefore, a positive value indicates that the model correctly assigned lower surprisal to the target verb in the plausible context relative to the implausible context, i.e., role-sensitivity, while a value close to zero or negative indicates that the model incorrectly assigned similar or greater surprisal to the verb in the plausible context than the implausible context.

6.5.2 Results

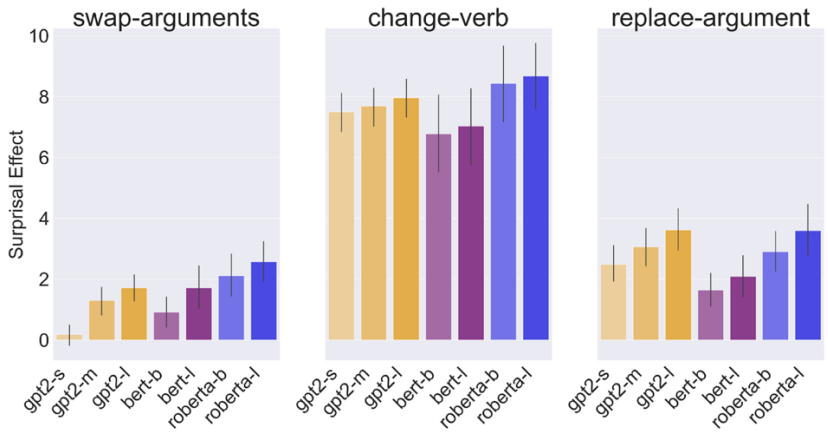


Figure 36: Surprisal effects plotted by condition and model. Higher values indicate greater role-sensitivity.

We report the surprisal effect in all the models in Figure 36. In line with our expectations, the surprisal effect is larger for the replace-argument items than the swap-arguments items, showing that models are less sensitive to role reversals compared to replace-arguments. GPT2-small in particular did not exhibit any sensitivity to the role-reversed sentences, while showing considerably more sensitivity to the replace-argument sentences, consistent with Chow et al. (2016). However, one key difference between the model and human responses is that all the models' effects for change-verb were far higher than both the swap-arguments and the baseline replace-argument case. Instead of showing a smaller effect, like for swap-arguments, the surprisal effect for these sentences is far higher.

The performance of GPT2-small for the swap-arguments condition mirrors the early stages of human processing more closely, as these role-reversed sentences do not elicit an N400 potential. However, humans are also not sensitive to the manipulation in the change-verb stimuli since they use an abstract, generalized representation of argument roles, which is a major contrast with the models' surprisal. Based on the comparably better performance on the change-verb and replace-argument conditions relative to swap-arguments, it is likely that the models are making use of

specific lexical cues to make their inferences rather than the structural relations humans are using. This is because the two conditions the model does better on introduce lexical variation in the stimuli, which is not the case for swap-arguments.

6.6 Experiment 6-2: Probing

6.6.1 Methods

While the surprisal estimates in Experiment 6-1 are computed based on the final layer of the models, in Experiment 6-2, we investigate which layers encode argument role information in verb representations by conducting a probing analysis. To show role-sensitivity at the verb, the model must correctly analyze the position of the arguments, represent the arguments with a role-specific meaning, and use that information to determine the plausibility of the verb that appears following the arguments. As these computations involve both syntactic and semantic processing, it is possible that such knowledge is encoded in earlier layers of the models which are not detectable in surprisal estimates based on final layer representations (Jawahar et al., 2019; Tenney et al., 2019). We investigate this by implementing layer-wise *probing classifiers* (Belinkov, 2022), on GPT2-small, which showed the most human-like pattern in the surprisal analysis, as well as GPT2-medium, BERT-large, and RoBERTa-large, which have the same number of layers and show better performance with the swap-arguments condition than GPT2-small.

For each condition, and for each layer, we train a logistic regression classifier on the models' representations of the target verbs, which predicts whether the verb is contextually appropriate or inappropriate. We choose to use a linear classifier because evidence points to conceptually relevant information being linearly separable in embedding space (Nanda et al., 2023). Target verbs in the plausible sentence were coded as 0 and the same target verbs in the implausible sentence were coded as 1.

Verb representations from each layer of each model were extracted using the minicons library. We report accuracies of each probe using 10-fold cross-validation with the scikit-learn implementation (Pedregosa et al., 2011). During training, we used a controlled method of splitting the train and test data sets, where the plausible and implausible verb pairs were always included in the same data set. This was to prevent the model from simply matching a verb in one context to the same verb in the counterpart context.

A high **classification accuracy** indicates that the verb representations extracted from the model contains information about the plausibility of the verb given the sentence it appears in - the model is able to distinguish contextually appropriate and inappropriate verbs.

6.6.2 Results

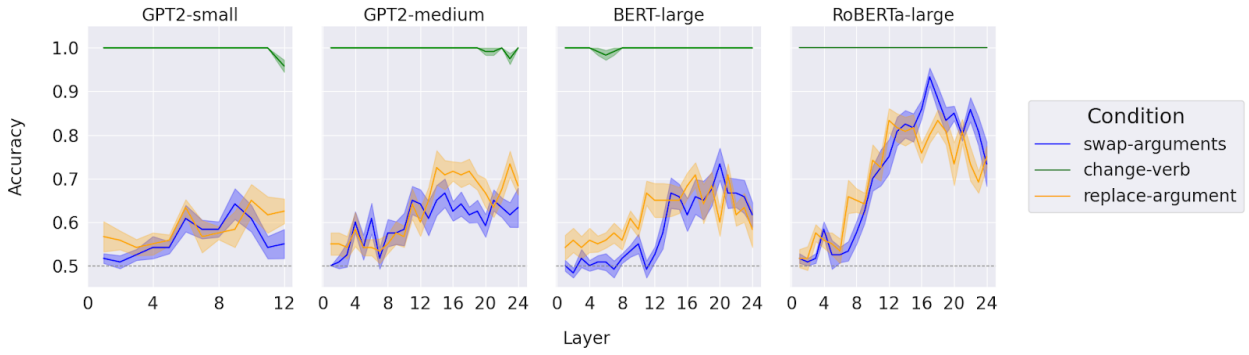


Figure 37: Classification accuracies for probes trained to distinguish plausible and implausible verbs under different conditions. Highlighted areas indicate standard errors of the mean across the 10 cross-validation folds. Dotted lines indicate at-chance accuracy.

The probes trained on the verb representations in the change-verb condition performed at ceiling for all models (Figure 37). This suggests that in all models, the systematic change in verb form (*-ed* vs. *-ing*) is robustly encoded in verb representations. This pattern corroborates the surprisal results, where the change-verb condition showed significantly large surprisal effects in all models, suggesting that the models can effectively distinguish verbs in the plausible and implausible contexts when the verb form differs between the two contexts.

Classification accuracy was generally lower for the conditions where the verb was kept the same and plausibility was determined by changing properties of the preceding context, i.e., swap-arguments & replace-argument, rather than verb form. GPT2-small did not improve greatly from chance-level performance. The larger models reached higher classification accuracy, with GPT2-medium and BERT-large reaching 70% accuracy, while RoBERTa showed the highest performance, reaching near 80-90% accuracy. For these larger models, decoding accuracy gradually increased throughout the layers and the particular increase in the middle layers suggests that verb plausibility information is more effectively represented from the middle layers.

While the accuracies between the swap-arguments and replace-argument conditions were overall comparable, the replace-argument condition showed slightly higher accuracy than the swap-arguments condition in earlier layers of BERT and RoBERTa, while the same contrast appeared in later layers of GPT-2 (small and medium). This suggests that role-dependent verb plausibility information may be encoded at different stages of processing in uni- and bi-directional models. Finally, there was a tendency for the accuracies to fluctuate more and even decrease at the final layers, particularly for the swap-arguments condition in RoBERTa, which drops from 90% to 70% accuracy. This suggests that role-dependent plausibility information may become partially lost in models' representations.

6.7 Experiment 6-3: Attention

6.7.1 Methods

One question based on the previous experiment findings is what gives rise to models' relatively weak performance on determining verb plausibility based on argument role information, particularly when the argument role is manipulated by swapping the position of the arguments (swap-arguments condition). One possibility is that for these items, the models often incorrectly

parse the argument roles indicated by the structure. It is possible that the models get confused about which noun is in which position and takes on which argument role. This could also offer a reason for why models perform better with the change-verb items, where argument position is fixed and held constant between the plausible and implausible conditions. In Experiment 6-3, we examine how models treat the preceding arguments by conducting an attention analysis that focuses on whether the models correctly allocate attention to the target subject at the verb position.

We adapt a similar method to that used in previous work. Ryu and Lewis (2021) inspected the attention patterns of GPT-2 in order to probe whether the presence of a partially-matching distractor word interferes with the model's processing of a subject-verb dependency. The authors found an attention head that was specialized in finding the subject and examined whether the attention to the target subject differed between the intervening and non-intervening conditions.

We compare the attention profiles of GPT2-small and RoBERTa-large, the models that performed the worst and best, respectively, in the previous experiments. For each model, we first define an attention head that allocates the greatest attention weight from the verb to the subject in the sentence. For example, given the sentence, *The restaurant owner forgot which customer the waitress served during dinner yesterday*, we calculated the attention weight from the verb *served* to the subject *waitress* for each layer and head. We define the attention head that had the greatest attention weight to the subject as the subject attention head. The selected subject attention head was then used to calculate the attention from the verb to the subject and object, respectively. A high attention weight to the subject and a low attention weight to the object indicate that the model correctly distinguishes subjects from objects.

Table 6: Results of the attention analysis. The values represent the subject attention head's average attention from the verb to the subject and its attention from the verb to the object under each condition. Standard deviations are in parentheses.

Model	Condition	Attention to Subject		Attention to Object	
		Plausible	Implausible	Plausible	Implausible
GPT2-small	swap-arguments	.53 (.15)	.53 (.17)	.18 (.10)	.19 (.06)
GPT2-small	replace-argument	.51 (.12)	.50 (.13)	.19 (.09)	.21 (.08)
RoBERTa-large	swap-arguments	.68 (.18)	.70 (.20)	.06 (.10)	.05 (.09)
RoBERTa-large	replace-argument	.65 (.16)	.68 (.16)	.06 (.08)	.04 (.02)

6.7.2 Results

For GPT2-small, we identified layer 3 head 10 (head indices: 2, 9) as the subject attention head, and for RoBERTa-large, we identified layer 13 head 16 (head indicies: 12, 15) as the subject attention head. The attention weight to the subject averaged across all items was .52 for GPT2-small and .68 for RoBERTa, indicating that these attention heads allocated most of the attention from the verb to the subject across the experiment items.

The results are shown in Table 6. We found similar attention patterns between the swap-arguments and replace-argument conditions. For both GPT-2 and RoBERTa, the subject attention head correctly allocates most of its attention to the subject rather than the object. However, RoBERTa gives less attention overall to the object than GPT-2 does, with the attention weight to the object remaining below 10%.

The results show that even GPT2-small, which did not show clear sensitivity to argument roles in the surprisal and probing analyses, correctly allocates attention to subjects with the subject head, though its attention is also distributed to the object more than the better performing RoBERTa-large. The attention analysis, therefore, suggests that it is unlikely that weak role-sensitivity at the verb arises from being confused about which argument is in which position or

which argument is assigned which role. Rather, the weak performance could be due to how the models encode the preceding argument role information into the representations of the verb. Models may be able to correctly distinguish argument roles but less capable of using this information to represent role-compatible and role-inappropriate verbs in different ways.

6.8 Discussion

While previous studies have examined language models' knowledge of argument roles by testing their capacity to distinguish plausible and implausible sentences, we take a new approach by examining whether models' representations of verbs in sentences encode plausibility based on preceding argument role information. This method, in combination with the controlled sets of materials used in psycholinguistic studies that examine human comprehenders' role-sensitivity, offers a rigorous and systematic test of language models' sensitivity to argument roles and a way to directly compare human and model behavior. In the surprisal and probing analyses, we find that language models generally exhibit greater sensitivity to changes in argument meanings than to changes in argument roles, similar to humans' initial predictions. However, unlike humans, they fail to show the same pattern across different types of argument role manipulations. Whether the argument role and verb compatibility is manipulated by swapping the argument positions or by changing the verb form, humans show the same processing pattern, whereas language models treat the two cases differently.

The relatively weak sensitivity to verb plausibility when the preceding arguments are swapped, which we observed in Experiments 6-1 and 6-2, is unlikely due to a misrepresentation of the context, as the models' attention patterns in Experiment 6-3 suggest that roles are accurately represented. Rather, we suggest it arises from the difficulty in evaluating whether a verb is plausible given the particular argument-role bindings enforced by the preceding context. This

involves a more complex analysis than simply computing context-independent argument and verb co-occurrences, which is potentially why humans' predictions fail to make use of such information rapidly during real-time prediction (Chow et al., 2016).

A key divergence between the model and human behaviors was with regard to which conditions caused more difficulty than others. Human comprehenders show the same pattern in the swap-arguments and change-verb conditions (i.e., no immediate role-sensitivity), both of which involve determining a verb's fit with respect to given argument roles. In all the models we tested, we observed greater performance in the change-verb condition than the swap-arguments condition. This suggests that language models treat the two conditions differently, diverging from human processing behavior. The contrast between the role-related conditions further indicates that models do not compute argument-verb relations in those contexts using a shared underlying process, unlike human comprehenders who show similar role-sensitivity regardless of whether verb plausibility is manipulated through swapping the argument roles or changing the verb aspect. A possible explanation for this divergence between models and humans is that different morphological inflections of the same root could be represented as separate items in the language models' vocabulary (e.g., devouring - devoured), as opposed to how humans process variations in verb aspect. These results indicate that language models, like humans, may show differences in responses to plausible and implausible words or sentences, but the specific conditions under which these contrasts emerge can diverge (also see Arehalli et al., 2022; Huang et al., 2024). This suggests that their performance may not rely on the same processing mechanisms as humans.

One notable observation was that GPT2-small showed stronger correspondence with the human N400 data patterns, while larger models showed the higher performance in all experiments, which outperformed humans' initial predictive processing capacities. GPT-2 and variants have

shown to be more effective at predicting human behavior compared to larger autoregressive models (Kuribayashi et al., 2023; Oh et al., 2023). Steuer et al. (2023) find a similar pattern, where smaller models predict human reading times better than larger ones that do better on syntactic and semantic judgments. Our results suggest that smaller models capture more immediate, online processing profiles of humans, and resemble human N400 patterns which reflect initial stages of predictive processing. Conversely, the measures derived from larger models more closely pattern with offline, final interpretations of humans. Nevertheless, no models capture the consistency between the two argument role manipulations which has been found with humans. These results offer insights into drawing connections with human empirical findings, especially for psycholinguists aiming to use language models, with regard to determining which models to use when simulating experiments. Additionally, the improved performance of larger models raises the question of whether scale is sufficient to learn these complex role-specific relationships; evaluating the argument role-reversal and replace-argument contrast for larger models like LLaMa (Touvron et al., 2023), as well as tracing the ability based on the number of parameters of a language model, e.g., the Pythia family of models (Biderman et al., 2023), can facilitate these types of investigations.

Our work provides a critical perspective to language models' representations of argument roles from a psycholinguistic perspective. Future directions could involve applying causal interpretability methods (Arora et al., 2024; Meng et al., 2022) to these sets of sentences. It may be the case that larger-scale models that assign correct plausibility ratings are implementing the similar computations for replace-argument and reversal items, which will take us further towards determining whether linguistic knowledge in language models is as robust as it seems.

6.9 Limitations

6.9.1 Cross-Linguistic Coverage

Our investigation was focused on English, but the role reversal effect has also been shown in languages like Mandarin (Chow & Phillips, 2018) and German (Ston et al., 2021). Although it is linguistically robust across humans, Xu et al. (2023) found that language model surprisal exhibits different trends in each of these three languages. Testing whether similar effects appear in other language models as well as monolingual or multilingual language models could be a way to establish whether the models' inferences are are based on language-specific factors or whether generalized representation of argument roles is an emergent phenomenon.

6.9.2 Interpretability

Although it is unclear the extent to which attention-based measures provide explanatory value for model outputs on a variety of NLP tasks, a review from Bibal et al. (2022) suggests that the use of attention to explain syntactic parses is promising. For our use case, attention heads that track dependencies are identified using correlational analyses, based on the weights between the verb and its arguments. A key future direction is to build on work in interpretability (Lakretz et al., 2021; Meng et al., 2022) which identifies causal mechanisms in language models responsible for specific computations. Arora et al. (2024) apply some of these measures to pairs of grammatical and ungrammatical sentences handling various syntactic phenomena. In future work, we hope to not just extend their methods, but derive measures of cognitive effort based on how the language models causally compute argument roles.

Chapter 7: Conclusion

This dissertation set out to investigate the cognitive mechanisms underlying anticipatory language processing, focusing on the role-reversal phenomenon as a case study of apparent failure of the system. While previous research has largely attributed role-reversal errors to faulty context-driven predictive processing, the precise source of these errors remained unclear. The primary objectives of this dissertation were threefold: (i) to decompose context-driven generation and integration into sub-processes, (ii) to determine which of these underlying processes are robust and which are more prone to errors, and (iii) to clarify the nature of the mechanisms such that role-sensitive expectations vary across experimental measures, sentence structures, and populations.

To address these questions, I introduced a framework for context-driven generation and integration (Figure 1, repeated from Chapter 1), outlining the key steps involved in real-time representation building and updating. While prior theories have often attributed role-reversal errors to early-stage, context-based generation failures, such as misrepresenting argument roles (Step 1) or failing to generate role-appropriate candidates (Step 2), the findings presented in this dissertation suggest a different perspective. Using multiple experimental and computational approaches, I demonstrated that early-stage processing may be more robust than considered previously and that the primary source of vulnerability arises when new input is encountered and integrated into the unfolding representation (Step 3).

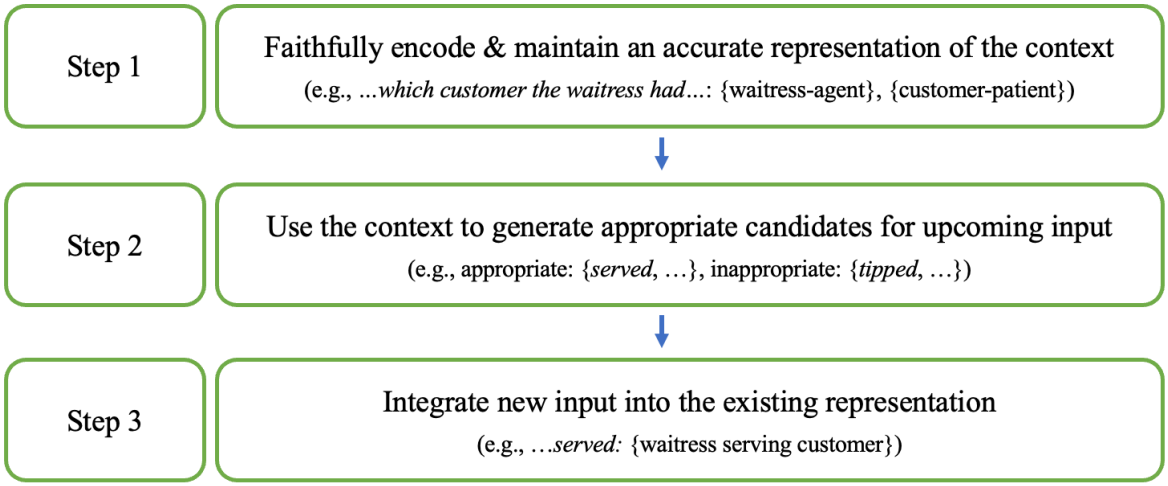


Figure 1: A schematic of the context-driven generation and integration process.

In Chapters 2 and 3, I demonstrated the robustness of generating anticipatory representations based on context, as well as the vulnerability in processes that occur after a word is encountered in a context. Specifically, the empirical results suggest that the vulnerability to role-reversals in comprehension may be less due to faulty generation of expectations and more due to the effects of encountering the role-reversal verb, when a strong candidate is not readily available given the preceding context. The carefully designed experimental paradigms helped to tease apart different sub-processes and to identify which parts are more vulnerable than others. In Chapter 2, a hybrid task was designed to narrow the gap between comprehension and production findings, placing participants in a setting where they had to be ready to generate a sentence continuation at any moment. Under these conditions, role-sensitivity was observed in both N400 responses and speeded cloze responses, demonstrating that people can immediately use argument roles to constrain their expectations, when they are required to quickly generate and commit to a single representation. These results suggest that Steps 1 and 2, specifically, parsing argument roles and generating role-appropriate verb candidates, are largely robust.

Chapter 3 provided additional evidence that the primary vulnerability to role-reversal errors lies in Step 3, particularly when comprehenders encounter a role-inappropriate verb before strong candidates have been generated. Item-wise analyses revealed that the speed with which comprehenders generated a strong alternative candidate was closely linked to their susceptibility to role-reversal lures. These findings suggest that the process of generating predictions based on context (Step 2) is more robust than previously assumed. Instead, much of the variability in role-reversal errors appears to arise during the integration of new input, especially when strong expectations have not yet formed. This highlights a critical point of vulnerability: when comprehenders are confronted with a verb that is only partially, but not fully, supported by the preceding context.

The empirical and modeling evidence presented in Chapter 4 sheds light on the nature of context-driven candidate generation (Step 2), demonstrating that it operates as a competitive race, where multiple candidates accumulate activation and compete to reach a selection threshold (Staub et al., 2015). Data from both experimental and simulation studies, specifically, speeded cloze tasks with adults and children, supported a central assumption of the race model: each candidate's strength is determined by its activation speed, and the relative speed of competing candidates determines the outcome, which candidate is selected and with what likelihood. Critically, the results showed that children and adults exhibited similar patterns only when timing, rather than probability, was used to index candidate strength. This finding supports the view that next-word generation is governed by a race-like process, and that this mechanism emerges early in development.

Chapter 5 showed the conditions under which the generation process is more likely to produce role-reversal candidates, in both children and adults. The key empirical finding was that

providing argument information (through visual contexts) before explicitly specifying thematic roles (through linguistic contexts) significantly reduced role-reversal generation in adults but not in children. This suggests that adults can begin generating predictions based on role-unspecified information (i.e., pictures of the arguments) and then quickly refine their expectations once thematic roles are assigned by the sentence context. This ability increases the likelihood that Step 2 will generate role-appropriate candidates. In contrast, children do not seem to benefit from the incremental introduction of argument information in the same way. Instead, they may wait to initiate candidate generation until the linguistic input is available. As a result, they show no significant difference between conditions where argument information is introduced gradually and those where some information about the arguments is made available prior to the full argument role context. This suggests that children may adopt a more conservative approach to using context to generate candidates, both in how early they begin the process and in how they integrate different sources of contextual information. The results from Chapter 5 further suggest that children may be in a vulnerable state longer than adults during comprehension, where they may be more situations where subsequent inputs are encountered when strong candidates have not yet been generated by the preceding context, making children particularly vulnerable to effects of confrontation observed with adults in Chapter 3. This remains to be examined in future work.

The comparison between human and artificial intelligence in Chapter 6 provided further evidence for the robustness of context-driven expectation generation processes (Steps 1-2) while suggesting that vulnerabilities stem from human-specific mechanisms. Unlike with humans, the distinction across different sub-processes is not existant in large language models—these are models trained to simply assign probabilities to the next word based on prior context. Humans, however, engage in different types of cognitive operations beyond next-word prediction, as

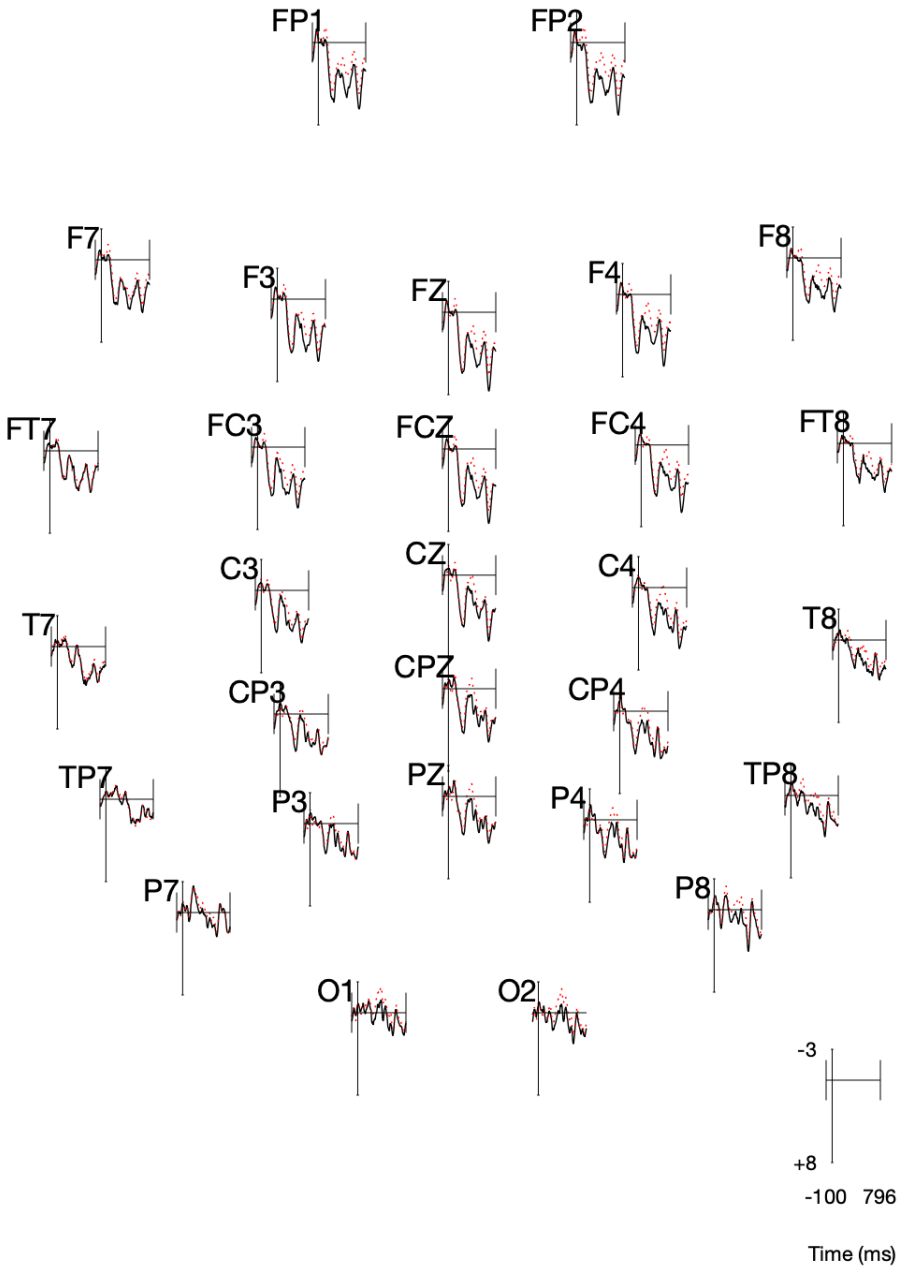
outlined in the proposed framework. If role-reversal errors were solely a consequence of data-driven predictive processing, similar to the way language models generate probabilities for upcoming words, then we would expect parallel patterns of errors in both humans and models. However, Chapter 6 revealed critical differences between human and model behavior, highlighting the limitations of explaining role-reversal effects purely through language-model-like probabilistic inference. Instead, the findings suggest that these errors arise from uniquely human mechanisms, including structured representation-building, candidate generation, and, crucially, the integration of new input into an existing representation. While large language models offer valuable insights into data-driven aspects of language prediction, they do not fully capture the fine-grained processes that introduce specific points of vulnerability in humans, as with role-reversal errors. In this sense, the findings from Chapter 6 higlight the limitation of using artifical intelligence systems as models of human language processing, while at the same time, they illustrate the benefit of using language models as a tool for identifying which aspects of human language processing can be attributed to probabilistic predictive processing and which require additional cognitive mechanisms beyond mere statistical next-word prediction.

To conclude, the findings of this dissertation refine our understanding of how the parser uses contextual information to generate expectations and integrate new information during real-time language processing. While anticipatory mechanisms enable efficient processing, the results suggest that apparent predictive failures, such as role-reversal errors, are not due to an inherent weakness in the representation-building system. Across different experimental paradigms, the results consistently demonstrate that context-based generation of anticipatory representations is largely robust and faithfully reflects preceding contextual information. The findings highlight that vulnerability arises at a rather specific point, when newly encountered input is integrated into an

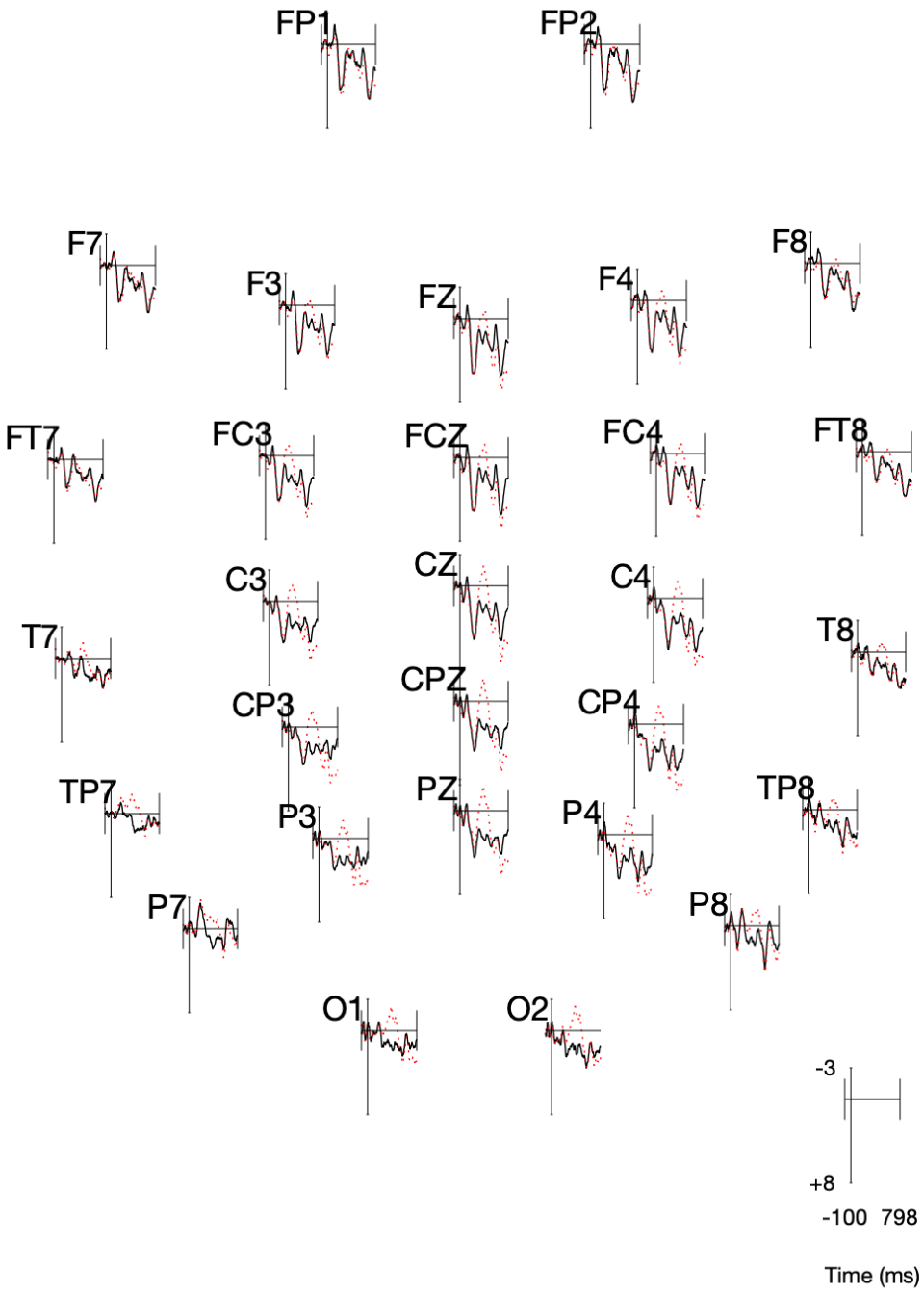
unfolding representation, and at a specific state, particularly when the generation process fails to produce a strong candidate before new information is introduced. These insights contribute to broader debates on the architecture of sentence processing, particularly regarding the distinction between constructing contextually constrained representations that align with one's grammar and the subsequent challenges of generating and integrating new information in real-time. By pinpointing where and why apparent failures occur, this dissertation not only refines existing models of language processing but also opens new avenues for research on how underlying processes are shared or diverge across different dimensions—between speaking and understanding, between children and adults, and between human cognition and artificial intelligence.

Appendices

A topographic display of Experiment 2-2 ERP results for critical items: Canonical (black) vs. Reversed (red).



A topographic display of Experiment 2-2 ERP results for control items: High-cloze (black) vs. Low-cloze (red).



Experimental items used in Experiment 4

Sentence fragment	**Example response**	**Normed cloze probability** (Block & Baldwin, 2010)
She could tell he was mad by the tone of his	voice	0.99
She went to the bakery for a loaf of	bread	0.98
The dentist says brushing your teeth twice a	day	0.98
After hitting the iceberg the ship began to	sink	0.96
Water and sunshine help plants	grow	0.95
She wore a colorful scarf around her	neck	0.95
When she got out of the car she closed the	door	0.95
The children went outside to	play	0.94
After dinner he washed his hands with	soap	0.94
I roasted the marshmallow over the	fire	0.89
The janitor accidentally spilled some water on the	floor	0.88
Charles dunked the basketball through the	hoop	0.87
At Jessie’s birthday party they ate a delicious	cake	0.95
There were no extra seats so she sat on the	floor	0.83
Derek's feet were cold, so he put on some	socks	0.82
After winning the carnival game, Tim received a	prize	0.82
The man happily sat down in the comfortable	chair	0.82

Maggie kept her wallet and keys inside her	purse	0.77
Pam did not have any clothes to	wear	0.74
The pinecone fell and hit Terry in the	head	0.74
She went to bed because she was	tired	0.72
On her birthday she excitedly opened the	gifts	0.71
The deer ran out of the woods and across the	road	0.67
Betty did not laugh when she heard the	joke	0.65
The child could not sleep without his stuffed	bear	0.63
She followed the recipe correctly to cook the	meal	0.56
They were startled by the sudden	noise	0.5
The milk sat out so long it began to	spoil	0.49
After being in the cold Jake's fingers were	numb	0.4
A large stone blocked the entrance to the	cave	0.36
During autumn the air is crisp and	cold	0.35
Jeffrey didn't get the question right but his answer was	good	0.23
Zack went to the supermarket and bought a	NA	NA
Bob wanted a snack so he ate a	NA	NA
Olivia opened the cabinet to take out the	NA	NA
Liam worked really hard to finish the	NA	NA
Emma peeled off the sticker and put it on the	NA	NA

Henry was holding the	NA	NA
Johnny was looking for his	NA	NA
Sophia was happy when she saw the	NA	NA

Appendix for Chapter 6

A. Computational Resources

Details of the model architectures we used are in Table 7. All experiments were run on a single CPU and took no more than two hours to run. We report metrics from a single run.

Table 7: Summary of model architectures. #L, #U, #H each refers to the number of layers, hidden units, and attention heads.

Model	Parameters	#L	#U	#H
GPT2 S	124M	12	768	12
GPT2 M	355M	24	1024	16
GPT2 L	774M	36	1280	20
BERT B	110M	12	768	12
BERT L	340M	24	1024	16
RoBERTa B	125M	12	768	12
RoBERTa L	355M	24	1024	16

B. Control Items

We further examined a set of items included in each study (Chow et al., 2016; Kim and Osterhout, 2005), where the plausibility of the verb was manipulated by simply replacing the target verb with another verb or associating the target verb with another argument. These materials have shown to elicit immediate neural responses in human comprehenders, indicating sensitivity to the likelihood of a target word appearing in a plausible context. High cloze conditions are listed first. a. Abby brushed her teeth after every meal/game and every snack. Chow et al. (2016). b. The [hungry boys]/[dusty tabletops] were devouring the plate of cookies when Jack arrived. Kim and Osterhout (2005), adapted. We computed the surprisal effect for plausible and implausible variants of the

same item for both studies, finding a much higher surprisal effect for both sets of control items (Figure 38) relative to the experimental conditions (Figure 36).

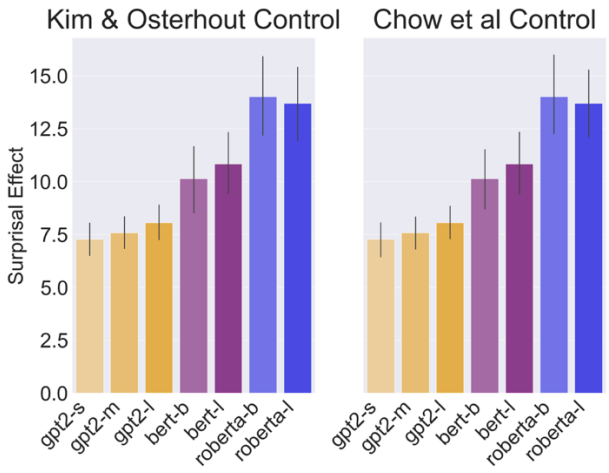


Figure 38: Surprisal effects for control items plotted by condition and model. Compare to change-verb for Kim & Osterhout, swap-arguments and replace-argument for Chow et al.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609.

Arora, A., Jurafsky, D., & Potts, C. (2024). CausalGym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*.

Aumeistere, A., Bultena, S., & Brouwer, S. (2022). Wisdom comes with age? The role of grammatical gender in predictive processing in Russian children and adults. *Applied Psycholinguistics*, *43*(4), 867-887.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics, 48*(1), 207–219.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., ... & Raff, E. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning* (pp. 2397–2430). PMLR.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436.

Brown, V. A., Fox, N. P., & Strand, J. F. (2022). "Where are the... Fixations?": Grammatical number cues guide anticipatory fixations to upcoming referents and reduce lexical competition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(5), 643.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language, 93*, 203–216.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143.

Burnsky, J. (2022). *What did you expect? An investigation of lexical preactivation in sentence processing* (Doctoral dissertation, University of Massachusetts Amherst). https://doi.org/10.7275/30435671

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review, 119*(2), 417.

Chow, W. Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience, 33*(7), 803–828.

Chow, W. Y., Smith, C., Lau, E., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience, 31*(5), 577–596.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*(4), 368–407. https://doi.org/10.1006/cogp.2001.0752

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120394.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547–619.

Ehrenhofer, L. (2018). *Argument roles in adult and child comprehension* (Unpublished doctoral dissertation). University of Maryland, College Park.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science, 7*(11), 180877.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505.

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology, 59*(1), e13940.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469–495.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*(2), 164–203. https://doi.org/10.1016/S0010-0285(03)00005-7

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*(4), 516–547. https://doi.org/10.1006/jmla.2000.2728

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language, 95*, 200–214.

Gambi, C., Pickering, M. J., & Rabagliati, H. (2016). Beyond associations: Sensitivity to structure in pre-schoolers' linguistic predictions. *Cognition, 157*, 340–351.

Gambi, C., Gorrie, F., Pickering, M. J., & Rabagliati, H. (2018). The development of linguistic prediction: Predictions of sound and meaning in 2-to 5-year-olds. *Journal of Experimental Child Psychology, 173*, 351–370.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*(8), 684–691.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://doi.org/10.3115/1073336.1073357

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology, 110*, 70–104.

Hanna, M., Belinkov, Y., & Pezzelle, S. (2023). When language models fall in love: Animacy processing in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12120–12135). https://doi.org/10.18653/v1/2023.emnlp-main.744

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience, 24*(4), 878–887.

Hickok, G. (2012). *The myth of mirror neurons: The real neuroscience of communication and cognition.* W. W. Norton & Company.

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. *Quarterly Journal of Experimental Psychology, 69*(6), 1056–1063.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*(1), 59–73.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language, 137*, 104510.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition, 224*, 105050.

Jacobs, C. L., Hubbard, R. J., & Federmeier, K. D. (2022). Masked language models directly encode linguistic uncertainty. In *Proceedings of the Society for Computation in Linguistics 2022* (pp. 225–228).

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., ... & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science, 47*(11), e13386.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52*(2), 205–225.

Kolk, H. H. J., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language, 85*(1), 1–36.

Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*(1), 23–42. https://doi.org/10.1016/j.cognition.2010.12.002

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience, 31*(5), 602–616.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research, 17*(1), 117–129.

Kuribayashi, T., Oseki, Y., & Baldwin, T. (2023). Psychometric predictive power of large language models. *arXiv Preprint arXiv:2311.07484*.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463–470.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163. https://doi.org/10.1038/307161a0

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition, 213*, 104699.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502.

Lee, E., Howitt, K., Dixon, L., Ness, T., Nakamura, M., & Phillips, C. (2023). Alignment between adult and child predictive processing dynamics: Evidence from a gamified cloze study in a museum. Poster presented at *The 36th Annual Conference on Human Sentence Processing*, University of Pittsburgh, Pennsylvania, USA.

Lelonkiewicz, J. R., Rabagliati, H., & Pickering, M. J. (2021). The role of language production in making predictions during comprehension. *Quarterly Journal of Experimental Psychology, 74*(12), 2193–2209.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–38.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition, 233*, 105359.

Liao, C. H., Lau, E., & Chow, W. Y. (2022). Towards a processing model for argument-verb computations in online sentence comprehension. *Journal of Memory and Language, 126*, 104350.

Lidz, J., White, A. S., & Baier, R. (2017). The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology, 97*, 62–78.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics, 7*, 195–212.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4*, 521–535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60.

Luck, S. J. (2022). *Applied event-related potential data analysis*. LibreTexts.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843–847.

Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science, 21*(7), 908–913.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific Reports, 8*(1), 1079.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

McCoy, T., Pavlick, E., & Linzen, T. (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*.

Meng, M., & Bader, M. (2021). The first noun as agent: Evidence from German passive sentence comprehension. *Journal of Psycholinguistic Research, 50*, 135–157. https://doi.org/10.1007/s10936-021-09765-y

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems, 35*, 17359–17372.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 1–29.

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv Preprint arXiv:2203.13112*.

Nair, S., & Resnik, P. (2023). Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Nakamura, M. (2023). *Generating and measuring predictions in language processing* (Doctoral dissertation, University of Maryland).

Nakamura, M., Momma, S., Sakai, H., & Phillips, C. (2024). Task and timing effects in argument role sensitivity: Evidence from production, EEG, and computational modeling. *Cognitive Science*. Advance online publication. https://doi.org/10.1111/cogs.70023

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (pp. 16–30).

Nation, K., Marshall, C. M., & Altmann, G. T. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology, 86*(4), 314–329.

Ness, T., & Meltzer-Asscher, A. (2021). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition, 207*, 104509.

Nieuwland, M. S., et al. (2018). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *PLOS Biology, 16*(5), e2004472. https://doi.org/10.1371/journal.pbio.2004472

Oh, B.-D., & Schuler, W. (2022). Entropy-and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9324–9334).

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics, 11*, 336–350.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Applebaum, S., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology, 6*, 384. https://doi.org/10.3389/fpsyg.2015.00384

Özge, D., Kornfilt, J., Maquate, K., Küntay, A. C., & Snedeker, J. (2022). German-speaking children use sentence-initial case marking for predictive language processing at age four. *Cognition, 221*, 104988.

Özge, D., Küntay, A., & Snedeker, J. (2019). Why wait for the verb? Turkish-speaking children use case markers for incremental language comprehension. *Cognition, 183*, 152–180.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics, 8*, 447–471.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002–1044.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modeling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour, 2*(9), 693–705.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Roland, D., Yun, H., Koenig, J. P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition, 122*(3), 267–279.

Roux, F., Armstrong, B. C., & Carreiras, M. (2017). Chronset: An automated tool for detecting speech onset. *Behavior Research Methods, 49*, 1864–1881.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences, 27*(11), 1032–1052.

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language, 29*(1), 86–102.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10), e2307876121.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Smolík, F., & Bláhová, V. (2022). Here come the nouns: Czech two-year-olds use verb number endings to predict sentence subjects. *Cognition, 219*, 104964.

Sommerfeld, L., Staudte, M., Mani, N., & Kray, J. (2023). Even young children make multiple predictions in the complex visual world. *Journal of Experimental Child Psychology, 235*, 105690.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17.

Steuer, J., Mosbach, M., & Klakow, D. (2023). Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 142–157).

Stone, K., & Rabovsky, M. (2025). The role of syntactic and semantic cues in preventing temporary illusions of plausibility. *Journal of Cognitive Neuroscience, 37*(4), 1–27. https://doi.org/10.1162/jocn_a_02320

Van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research, 22*(2), 241–255.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190.

Van Wonderen, E., & Nieuwland, M. S. (2023). Lexical prediction does not rationally adapt to prediction error: ERP evidence from pre-nominal articles. *Journal of Memory and Language, 132*, 104435.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392.

Wilcox, E. G., Futrell, R., & Levy, R. (2023). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–44.

Wilson, M., Petty, J., & Frank, R. (2023). How abstract is linguistic generalization in large language models? Experiments with argument structure. *Transactions of the Association for Computational Linguistics, 11*, 1377–1395.

Xiang, M., & Kuperberg, G. R. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience, 30*(6), 648–672.

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721).

Zehr, J., & Schwarz, F. (2018). PennController for internet-based experiments (IBEX). Retrieved from https://github.com/addrummond/ibex