Prediction and Memory Retrieval in Dependency Resolution

by

Tzu-Yun Tung

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in The University of Michigan
2024

Doctoral Committee:

Associate Professor Jonathan R. Brennan, Chair
Lecturer Lisa Levinson
Professor Richard L. Lewis
Professor Acrisio Pires

Tzu-Yun Tung

tytung@umich.edu

ORCID iD: 0000-0003-2447-6962

ACKNOWLEDGEMENTS

This dissertation characterizes my academic journey starting from the University of Michigan. It would not have been possible without all the wonderful people I met along the way.

First and foremost, my sincere gratitude goes to my advisor Jonathan R. Brennan. Jon is the mentor, researcher and colleague I can only aspire to be. He always encourages me to pursue my research interests, patiently listens to my half-baked ideas, and promptly offers essential insight into an actionable plan. Beside scientific rigor, curiosity, and professional development, he also genuinely cares for the well-being of his mentees. His door is always open, and he is always available to provide words of wisdom and support. Jon brings out the best potential in me. And for that, I am grateful.

I had the privilege to be formally introduced to cognitive science and the memory model of sentence processing by Richard L. Lewis. Rick is my role model for conceptual precision, thoughtful discussion, and unparalleled humor. He has the magic of elevating my projects to the next level by viewing them from multiple perspectives and asking crucial questions I have yet to consider. Working with him is fun, inspiring, and enlightening at every moment. I am beyond appreciative.

I am deeply indebted to Lisa Levinson for her expertise and insight in formal linguistic theories, experimental design, result interpretation, and the use of large language models. Lisa is the dream mentor who generously takes time to help her mentees think (and re-think) through their projects at every step of the development. All our conversations are filled with discovery, excitement and revelation.

I am also greatly thankful to Acrisio Pires for his invaluable feedback from theoretical linguistics. Acrisio carefully guided me to probe deeper into the mental representation of the linguistic knowledge and its interaction with the memory operations. He also

immensely sharpened my academic reasoning and writing.

I could not have better introduction to the psychology of language than by Julie Boland. Julie thoughtfully helped shape my qualified research paper from its conception, design, result interpretation to discussion. Words cannot express my gratitude.

In the Departments of Linguistics and Psychology, I am fortunate to learn from interactions with Natasha Abner, Steve Abney, Marlyse Baptista, Patrice Speeter Beddor, Andries W. Coetzee, San Duanmu, Nick Ellis, Samuel David Epstein, Benjamin Fortson, Susan Gelman, Ezra Keshet, Jelena Krivokapić, Savithry Namboodiripad, Robin Queen and Sarah Thomason. They showed me how to become an outstanding scholar in multiple ways.

I am also blessed to have numerous inspiring conversations with Emily Atkinson, Ina Bornkessel-Schlesewsky, Luca Campanelli, Chun-Yin Doris Chen, Zhong Chen, Chia-Ju Chou, Brian Dillon, Simon E. Fisher, Kyle Johnson, Alan Hezao Ke, Li-Chuan Ku, Dave Kush, Marco Chia-Ho Lai, Daniel Chi Dat Lam, Chia-Lin Lee, Chia-Ying Lee, Chia-Hsuan Liao, Chien-Jer Charles Lin, Keng-Yu Lin, Andrew McInnerney, Tamara Swaab, Xin Sun, Matthew W. Wagers and Ming Xiang. Thank you for discussing with me and stopping by my posters during conferences. You enriched my projects with so many brilliant ideas.

I would not have survived this journey without the emotional and professional support from all fellow members of the Computational Neurolinguistics Lab. Thanks go to David Abugaber, James Baybas, Lauretta Cheng, Jeonghwa Cho, Justin Craft, Haoyu Du, Samia Elahi, Tamarae Hildebrandt, Chia-wen Lo, Emily Sabo, Csilla Tatar, Rachel Weissler, Shuchen Wen and Junyuan Zhao. They have been with me through all my wildest projects.

My fellow colleagues and dear friends also made my time at Michigan joyful and memorable. Special thank you to Aliaksei Akimenka, Rawan Bonais, Dominique Bouavichith, Danielle Burgess, Dominique Canning, Yu-Chuan (Lucy) Chiang, Wilkinson Daniel Wong Gonzales, Jiseung Kim, Mathew Alex Kramer, Joy Peltier, Yourdanis Sedarous, Jungyun Seo, Yushi Sugimoto, Stephen Tobin, Kelly Wright and Jian Zhu. I will never forget hanging out at the grad lounge, celebrating holidays, and always having good laughs together. I would also like to thank all my friends in the US and Taiwan for their warm encouragement, companionship, and all sorts of support whenever I needed them.

Lastly, I would not be able to embark on this journey at the first place without the unwavering and unconditional support from all my family. Thank you for always believing in me, being there for me, and making this journey full of love and laughter. I am truly blessed.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Successful language comprehension requires the rapid deployment of working memory resources alongside the capacity to predict upcoming linguistic input. While previous research views these as competing factors, this dissertation explores a unified theory of processing complexity and evaluates the interaction between memory and prediction. The evaluation focuses on how language-users deploy these factors to form long-distance dependencies in Mandarin. Specifically, I investigate how memory retrieval of a target word is affected by: (i) the time elapsed since the word first appears, (ii) interference from a neighboring distractor word that shares some linguistic features, and (iii) linguistic expectations of the target word. Neuroelectric signals of the human brain during naturalistic language comprehension were acquired by two electroencephalography (EEG) experiments. The experiments examine the resolution of noun-phrase ellipsis and subject-verb agreement using, respectively, carefully designed experimental stimuli and a naturally occurring audiobook story. The data are analyzed in terms of their fit to the quantitative predictions from computational models of these expectation and memory retrieval processes. This approach allows for a comparison between predictions of a symbolic cognitive model and a non-symbolic large language model. I report the first ever empirical evidence of the modulation of memory retrieval by linguistic expectations with a controlled experiment. I then report the first cortical electrophysiological evidence of the memory effects during naturalistic story listening, and suggest that interference modeled with cue-based working-memory retrieval framework may generalize to more everyday comprehension situation. The primary contributions of this work are, first, to unveil the biological underpinning of cognitive operations essential for how people understand complex sentences in a way that generalizes across languages and second, to contribute to methodological

advancement in combining computational modeling and cognitive neuroscience to study naturalistic language comprehension in real time which can be generalize to real-world situations.

CHAPTER I

Introduction

Language is an integral part of what makes us human, allowing us to express an infinite set of meanings. As a highly efficient cognitive system, it computes the arbitrary mapping between form and meaning. In language comprehension, this algorithm transforms speech, sign or written input to concepts understood by language users with a remarkable speed of two to three English words per second in a deceivingly effortless fashion. The question of how the brain implements this powerful algorithm in real-time language processing remains at the core of cognitive science.

I draw on computational tools to model two key facets of this system and test models against neuroelectric brain signals recorded during language comprehension. These facets, drawn from current research, are: (i) linguistic predictions (Hale, 2001; Levy, 2008) and (ii) working memory load (Lewis et al., 2006; McElree, 2006; Vasishth et al., 2019). While these two facets of processing have received significant attention, relatively little research seeks to integrate them into a unified model (Vasishth and Drenhaus, 2011; Levy, 2013). My research pursues just such integration, guided by the following specific questions:

1. How do predictions and working memory demands influence word-by-word language processing?
2. What are the neural mechanisms that mediate these effects?
3. How might these factors individually or jointly explain patterns of comprehension across different languages?

Answers to these questions form the initial steps to revealing the underlying mechanisms that support language comprehension in a way that generalizes across the diversity of the world's ≈7,000 languages.

This dissertation investigates how readers establish a dependent relation between two linguistic elements far away from each other in a sentence. This "long-distance dependency" is crucial for understanding "who did what to whom" in a sentence, and can be used to test how readers make predictions and/or consume memory resources during the process. The modulation of probabilistic expectations on memory cost will be explored, alongside a nuanced perspective for memory load. In contrast to the heavily studied distance-based memory cost (Gibson, 2000; Hsiao and Gibson, 2003; Vasishth and Drenhaus, 2011) that focuses on the distance between two co-dependent linguistic items (i.e. locality), recent psycholinguistic research has turned to cost associated with the distinctiveness of items stored in memory according to the representational features of items (Vasishth et al., 2019). When retrieving a target linguistic element that occurs earlier in an utterance to establish a long-distance relationship between the current and the retrieved items, memory retrieval can be affected if an intervening lexical item (a "distractor") shares some linguistic features with the target; this is interference, which appears in multiple guises. In "inhibitory interference", a distractor may cause a slow-down and inaccuracy in memory retrieval of the target item, and thus jeopardize the comprehension of an otherwise grammatical sentence (Franck et al., 2015; Jäger et al., 2017; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Van Dyke and McElree, 2011). Intriguingly, "facilitatory interference" occurs in ungrammatical sentences, where distractors seemingly create an "illusion of grammaticality" and thus make processing easier (Cunnings and Sturt, 2018; Dillon et al., 2013; Jäger et al., 2017, 2020; Lago et al., 2015; Parker and Phillips, 2017; Sturt, 2003; Tucker et al., 2015; Wagers et al., 2009). In sum: Intervening distractors lead to more uniqueness-based memory cost in grammatical sentences, but less cost in ungrammatical situations. Therefore, interference has been linked to language comprehension difficulty, after long being considered a major contributor to forgetting in the domain general memory research (Van Dyke and McElree, 2011).

(1) provides an example set from Sturt (2003) following the notation convention of Engelmann et al. (2019). "The surgeon" is the binding accessible antecedent (i.e. the target) of the reflexive "himself", while "Jonathan" and "Jennifer" are the binding inaccessible antecedents (i.e. the distractors). When encountering the reflexive, two relevant retrieval cues are used to retrieve the target: c-command and the gender of reflexive, masculine. The former cue separates the target from the distractor, and the latter cue relates to condition manipulation. The match or mismatch with respect to individual retrieval cues is represented by the + and - in the feature matrix of the noun phrases. "The surgeon" is a stereotypically masculine noun phrase in the study. The target fully matches the retrieval cues in both a. and b. In contrast, the partial feature match of the distractor in a, but not b, is predicted to give rise to inhibitory interference.

(1) a. Target-match; Distractor-match (Interference)
The **surgeon**$^{+MASC}_{+CCOM}$ who treated **Jonathan**$^{+MASC}_{-CCOM}$ had pricked **himself**$\left\{\begin{smallmatrix}MASC\\ CCOM\end{smallmatrix}\right\}$
with a used syringe needle.

b. Target-match; Distractor-mismatch (No interference)
The **surgeon**$^{+MASC}_{+CCOM}$ who treated **Jennifer**$^{-MASC}_{-CCOM}$ had pricked **himself**$\left\{\begin{smallmatrix}MASC\\ CCOM\end{smallmatrix}\right\}$
with a used syringe needle.

The second type of interference emerges when both the target and distractor only partially match the retrieval cues. The overall speedup at the retrieval site in reading time marks it as facilitatory interference (Dillon et al., 2013; Engelmann et al., 2019; Logačev and Vasishth, 2016). An example set can be found in (2) (Sturt, 2003; Engelmann et al., 2019). The target "the surgeon" is stereotypically masculine. Therefore, it mismatches the gender retrieval cue, feminine, of the reflexive "herself" in both a. and b. The distractor "Jennifer" in a. is nevertheless +FEM, examplifying the facilitatory interference criteria.

(2) a. Target-mismatch; Distractor-match (Interference)
The **surgeon**$^{-FEM}_{+CCOM}$ who treated **Jennifer**$^{+FEM}_{-CCOM}$ had pricked **herself**$\left\{\begin{smallmatrix}FEM\\ CCOM\end{smallmatrix}\right\}$
with a used syringe needle.

b. Target-mismatch; Distractor-mismatch (No interference)
The **surgeon**$^{-FEM}_{+CCOM}$ who treated **Jonathan**$^{-FEM}_{-CCOM}$ had pricked **herself**$\left\{\begin{smallmatrix}FEM\\ CCOM\end{smallmatrix}\right\}$
with a used syringe needle.

Other than retrieval interference, the predictability of a word or structure given a context has also been associated with processing ease or difficulty in reading comprehension (Hale, 2001; Levy, 2008). Comprehenders make linguistic expectancies at various levels based on the sentential context (Kutas et al., 2014). Prediction of upcoming structure, for example, has been proposed to constrain possible syntactic violation and thus enable rapid syntactic analysis, showing effects as fast as 200 ms after stimulus onset during online word-by-word reading (Lau et al., 2006). Federmeier et al. (2007) focused on lexical prediction, reporting the beneficial effect of processing expected words from 300 to 500 ms post-stimulus-onset, and the costly effect of processing unexpected (but plausible) words in strongly, but not weakly, constraining contexts at the 500-900 ms time window. While the earlier stage was suggested to reflect facilitated processing due to higher degree of match to expectation, the later stage might reflect the recognition of mismatch and/or additional resources needed for the revision of expectation.

It remains unclear, nevertheless, when and how linguistic predictions may influence the memory effects when language users try to form a long-distance relation between two non-adjacent elements. Only one prior study addressed this question directly; they report an interaction between predictability and inhibitory interference, but rely on a "dual-task" (i.e. word recall and sentence-reading) in English that is quite different from every-day language comprehension (Campanelli et al., 2018). It is crucial to investigate whether those findings can be generalized across linguistic dependencies across languages and within a more natural comprehension paradigm. When the target item is highly anticipated in grammatical sentences, will it become more distinctive from the distractor, reducing the inhibitory interference effect? In contrast, when a highly predictable target causes a prediction error in ungrammatical sentences, will it become less distinctive from the distractor, increasing the facilitatory interference effect? The dissertation aims to first replicate the interference and expectation effects separately, and subsequently investigate the degree, and neural time course, of their interaction. I test this question by measuring the electrical activity of large sets of neurons using Electroencephalography (EEG); this neural signal indexes the processing of long-distance dependency in real time. I will first examine the electrophysioligical correlates of the memory and prediction mechanisms separately, and then use those neural signals to investigate how they interact. Importantly, I use carefully constructed experimental stimuli along-side an audiobook story, which are crucial for testing how well the models under consideration account for language processing not only in the artificial conditions of a laboratory, but in more every-day circumstances. Two separate EEG experiments will unveil the use of linguistic expectancies and working memory resources during language comprehension, and their relative contribution to processing complexity.

This dissertation will contribute to a unified and multidimensional theory of processing complexity by concurrently examining the role of linguistic prediction and memory interference during cross-linguistic online comprehension with high ecological validity. Those cognitive operations will be explicitly modeled by computational theories of comprehension, and grounded in the electrical dynamics of neural activity. The study will address the interplay of human memory and linguistic predictions, a key intersection that is underspecified in dominant frameworks. The study will also have theoretical implications for the functional interpretation of neural signals, shedding light on the cognitive processes and computations carried out in the electric activities of population of neurons.

To answer these questions in this dissertation, I will first present an EEG experiment using constructed stimuli of Mandarin noun-phrase ellipsis in Chapter II. The inclusion of both grammatical and ungrammatical conditions enables the examination of both inhibitory and facilitatory interference effects discussed in previous literature. The neural correlates (i.e., event-related potentials (ERPs)) of predictions and memory retrieval will also be introduced in the chapter. Afterwards, a second EEG experiment using a Mandarin audiobook story will be presented to evaluate the theories in a more natural and ecologically valid setting in Chapter III. I will demonstrate how to combine cognitive neuroscience and computational modeling to study naturalistic language comprehension in real time. Chapter IV will conclude the dissertation by considering the implications of the results obtained from the two EEG experiments, as well as providing several ways forward for the next exciting steps.

CHAPTER II

Expectations Modulate Retrieval Interference during Ellipsis Resolution

Abstract

Memory operations during language comprehension are subject to interference: retrieval is harder when items are linguistically similar to each other. We test how such interference effects might be modulated by linguistic expectations. Theories differ in how these factors might interact; we consider three possibilities: (i) predictability determines the need for retrieval, (ii) predictability affects cue-preference during retrieval, or (iii) word predictability moderates the effect of noise in memory during retrieval. We first demonstrate that expectations for a target word modulate retrieval interference in Mandarin noun-phrase ellipsis in an electroencephalography (EEG) experiment. This result obtains in globally ungrammatical sentences – termed "facilitatory interference." Such a pattern is inconsistent with theories that focus only on the need for retrieval. To tease apart cue-preferences from noisy-memory representations, we operationalize the latter using a Transformer neural network language model. Confronting the model with our stimuli reveals an interference effect, consistent with prior work, but that effect does not interact with predictability in contrast to human EEG results. Together, these data are most consistent with the hypothesis that the predictability of target items affects cue-preferences during retrieval.

Keywords: Predictability, Cue-based Retrieval, Mandarin, Sentence Processing, P600, GPT2.

2.1 Introduction

Successful language comprehension requires retrieving previously encountered linguistic items, and may be subject to interference from other recent or similar entities in memory (Cunnings and Sturt, 2018; Dillon et al., 2013; Franck et al., 2015; Jäger et al., 2017, 2020; Lago et al., 2015; Martin et al., 2012; Martin, 2018; Sturt, 2003; Tucker et al., 2015; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Van Dyke and McElree, 2011; Wagers et al., 2009; Vasishth et al., 2019). Recent research now hints at an interaction between retrieval interference and predictive processing (Campanelli et al., 2018; Futrell et al., 2020; Parker and Phillips, 2017; Schoknecht et al., 2022; Tanner et al., 2014). One thing that remains to be established is how expectations of the target item might modulate the interference effect under both successful and unsuccessful retrievals. In example (1), readers must retrieve the encoding of **shirt** from memory upon reaching **one**, in order to interpret “one” as “one shirt”, not “one book”.

(1) Anne brought a **shirt** that was next to the **book** and Emma also brought **one**.

The retrieval process is guided by cues provided at the retrieval site, including structural characteristics which point to “shirt” as the target antecedent (e.g., see Dillon et al., 2013; Kush et al., 2017, for using syntactic information to constrain antecedent retrieval). Specifically, the two conjuncts in this coordinate sentence requires symmetrical or parallel syntactic and semantic representations; broad sources of evidence for this symmetry are reviewed by Zhang (2009). In the present example, this requirement renders “shirt”, but not “book”, as the target antecedent. If the retrieval cues fail to uniquely map onto the target element in memory, interference from other cue-matching distractor elements may surface.[1]

Mandarin Chinese presents a useful testing ground for interference effects because

[1] The terms “distractor” and “attractor” are used interchangeably in this study.

the retrieval site offers both semantic and structural cues in the noun-phrase ellipsis construction; this enables examination of interference from semantic-cue-matching distractors which are structurally incorrect as antecedents. As detailed below, Mandarin classifiers at a retrieval site provide both syntactic and semantic cues that can map onto the target antecedent and/or the intervening distractor. The present study leverages this property to investigate how cue-based retrieval might interact with probabilistic expectations during sentence comprehension, as indexed by event-related potentials (ERPs).

2.1.1 Cue-based retrieval

The cue-based retrieval framework theorizes that the retrieval of target lexical items from memory depends on the match between memory contents and retrieval cues (Lewis and Vasishth, 2005; Lewis et al., 2006; McElree, 2000; McElree et al., 2003; Van Dyke and Lewis, 2003). Lexical items are stored with features of their intrinsic (e.g., lexical and morphological) properties along with features encoding the local syntactic context they appear in. Re-activation of an item is contingent on the degree of match between those features and retrieval cues provided at the retrieval site. As a consequence of this architecture, distractor items matching a subset of those retrieval cues may lead to two types of similarity-based interference.

In grammatical sentences, "cue overload" may occur when a syntactically licensed target word fully matches the cues while a syntactically unlicensed distractor partially matches them. This results in inhibitory interference which leads to slower reading time at the retrieval site in self-paced and eye-tracking studies (Franck et al., 2015; Jäger et al., 2017; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Van Dyke and McElree, 2011).

To illustrate inhibitory interference, example (2) provides a set of sentences with subject-verb dependencies (Dillon et al., 2013). Here "The new executive" is the target subject noun of the verb "was", while "manager" and "managers" are distractor nouns. When encountering the verb, two relevant retrieval cues are used to retrieve the target: structural location in terms of Local Subject-hood and Singular verbal number. The former syntactic cue distinguishes the target from the distractor while the latter cue is manipulated in this study. The match or mismatch with respect to individual retrieval cues is represented by the + and − in the feature matrix of the noun phrases, following the convention of Engelmann et al. (2019) and Jäger et al. (2020). While the number feature matches the distractor in the (a) sentence, it does not in (b); the cue overload in (2) (a) causes inhibitory interference.

(2) a. Grammatical; Interference
The new $\textbf{executive}^{+Singular}_{+LocalSubject}$ who oversaw the middle $\textbf{manager}^{+Singular}_{-LocalSubject}$ apparently $\underline{\textbf{was}}\left\{\begin{smallmatrix} Singular \\ LocalSubject \end{smallmatrix}\right\}$ dishonest about the company's profits.

b. Grammatical; No Interference
The new $\textbf{executive}^{+Singular}_{+LocalSubject}$ who oversaw the middle $\textbf{managers}^{-Singular}_{-LocalSubject}$ apparently $\underline{\textbf{was}}\left\{\begin{smallmatrix} Singular \\ LocalSubject \end{smallmatrix}\right\}$ dishonest about the company's profits.

A second type of interference emerges in ungrammatical sentences when both the target and distractor only partially match the retrieval cues. Such cases show evidence of a processing speedup at the retrieval site, termed "facilitatory interference" (Cunnings and Sturt, 2018; Dillon et al., 2013; Jäger et al., 2017, 2020; Lago et al., 2015; Parker and Phillips, 2017; Sturt, 2003; Tucker et al., 2015; Wagers et al., 2009). An example is shown in (3) (Dillon et al., 2013; Engelmann et al., 2019; Jäger et al., 2020). Here, the target "the new executive" is singular and thus it fails to match the number retrieval cue of the plural verb "were" in both (3) (a) and (b). The distractor "managers" in (3) (a) is +Plural, matching that retrieval cue, with the consequence that retrieval might be "successful" (if incorrect) even in ungrammatical contexts. This is facilitatory interference. Interestingly, not all interference effects have the same strength, which we turn to in the next section.

(3) a. Ungrammatical; Interference
*The new $\textbf{executive}^{-Plural}_{+LocalSubject}$ who oversaw the middle $\textbf{managers}^{+Plural}_{-LocalSubject}$ apparently $\underline{\textbf{were}}\left\{\begin{smallmatrix} Plural \\ LocalSubject \end{smallmatrix}\right\}$ dishonest about the company's profits.

b. Ungrammatical; No Interference
*The new $\textbf{executive}^{-Plural}_{+LocalSubject}$ who oversaw the middle $\textbf{manager}^{-Plural}_{-LocalSubject}$ apparently $\underline{\textbf{were}}\left\{\begin{smallmatrix} Plural \\ LocalSubject \end{smallmatrix}\right\}$ dishonest about the company's profits.

2.1.2 Interference asymmetry and prediction

Agreement attraction errors seem to vary according to both grammaticality and dependency types. Previous studies have noted two distinct kinds of processing asymmetries. First of all, facilitatory interference effects in ungrammatical sentences appear to be stronger than the inhibitory interference observed in grammatical sentences. This "grammatical asymmetry" was first reported in Wagers et al. (2009), who investigated English subject-verb agreement in a self-paced reading paradigm; they found faster reading time in the presence of a number-matching attractor noun, compared to a number-mismatching attractor, in ungrammatical, but not grammatical, conditions. In that study, attraction from a prepositional modifier ameliorates the effect of subject-verb agreement violation (i.e. shows facilitatory interference), but does not affect grammatical sentences (i.e. no inhibitory interference). This grammatical asymmetry is further supported by a Bayesian random-effects meta-analysis of 77 experimental comparisons from eye-tracking and self-paced reading studies (Jäger et al., 2017), although not entirely supported by a recent forced-choice judgment study (Hammerly et al., 2019), which highlights the importance of further investigation.

Using EEG, Tanner et al. (2014) did replicate this "asymmetrical attraction effect" with English subject-verb agreement. Tanner et al. attribute this asymmetry to a predictive mechanism which operates alongside retrieval in the following way. For obligatory constituents like the upcoming verb for an English subject noun, the syntactic structure and the specifications for verbal agreement features are automatically predicted. In grammatical sentences, successful prediction minimizes the need for further retrieval (Dillon et al., 2013; Lago et al., 2015; Wagers et al., 2009), and also reduces the occurrence of attraction effects, which happen during the retrieval process. But in ungrammatical sentences, the actual bottom-up input (with mismatched verbal agreement features) clashes with the top-down predictions. This mismatch triggers retrieval for the (wrongly) predicted features, and gives attractors a chance to cause attraction during the retrieval process. By this logic, the retrieval process and attraction effects are biased toward ungrammatical, not grammatical sentences.

This account also captures the apparent symmetry in attraction effects observed for grammatical and ungrammatical Spanish noun-phrase ellipsis by Martin et al. (2012). Tanner et al. (2014) reason that since no reliable prediction can be made for the computation of noun-phrase ellipsis in those stimuli, retrieval needs to be initiated for both grammatical and ungrammatical sentences, giving rise to attraction symmetry. In sum, Tanner et al. propose that the need for retrieval relies on the predictability of linguistic dependencies. For predictable dependencies such as subject-verb agreement, retrieval and attraction occur when prediction is unsuccessful (as in ungrammatical sentences), but not when prediction is successful (as in grammatical sentences). For unpredictable dependencies such as noun-phrase ellipsis, retrieval and attraction occur for both grammatical and ungrammatical sentences because no prediction is available. We will call this model the "retrieval-by-predictability" account to highlight its unique features in comparison to alternatives, discussed below.

Recently, predictive models based on noisy memory representation have also been proposed to account for the "grammatical asymmetry" observed in human data. Both lossy-context surprisal model (Futrell et al., 2020) and Transformer-based neural network language models such as Generative Pre-trained Transformer-2 (GPT2) (Radford et al., 2019) use noisy memory representations of the previous context to calculate predictability of the next word. The noisy memory representation is a version of the true context with some information obscured by noise. The incomplete information about previous words reflects the general information-loss characteristic of of memory representations. Hahn et al. (2022) further proposes that the memory representations are noisy because they are refined to reduce processing cost due to cognitive resource constraints, which the authors call the resource-rational model of fine-grained memory representations.

Word predictability, formalized as surprisal values obtained from the predictive models based on noisy memory representation, have been associated with processing difficulty during comprehension (Futrell et al., 2020). When cast in terms of surprisal values from a Transformer-based artificial neural network, like GPT2, Ryu and Lewis (2021) demonstrate that such an account successfully simulates the presence of facilitatory interference effects in ungrammatical sentences as well as the absence of inhibitory interference effects in grammatical sentences of English subject-verb agreement and reflexive-antecedent dependencies. In brief, Ryu and Lewis argue that word predictability based on noisy memory representations may directly characterize the interference profile during retrieval. We will call this theoretical alternative the "noisy-memory-based-predictability" account.

A second processing asymmetry concerns facilitatory interference effects in ungrammatical sentences: such effects seem to be stronger in subject-verb agreement dependency than in reflexive-antecedent dependency. This "type asymmetry" was first documented in Dillon et al. (2013) in their English eye-tracking experiments, and later reinforced by the large scale Bayesian random-effects meta-analysis of 77 experimental comparisons from Jäger et al. (2017). More empirical work is needed as a recent large-sample replication study by Jäger et al. (2020) did not find type asymmetry with eye-tracking measures.

To account for type asymmetry, Parker and Phillips (2017) also appeal to the predictability of linguistic dependencies, which they propose affects cue weightings during the retrieval process. For unpredictable dependencies such as holds between a reflexive and its antecedent, retrieval is part of the necessary resolution process and structural cues are prioritized, minimizing facilitatory interference effects. That is, the antecedent does not serve to predict any upcoming reflexive anaphor in ungrammatical reflexive-antecedent dependency. The default retrieval mechanism thus ensures priority for structural cues in such cases.

In contrast, for predictable dependencies like subject-verb agreement, retrieval is triggered by prediction error and thus structural cues are not prioritized, resulting in facilitatory interference effects. Put simply, the subject noun predicts the number of the verb. If a number prediction error occurs in ungrammatical subject-verb agreement, retrieval is invoked as a repair mechanism (Lago et al., 2015; Wagers et al., 2009). Due to this prediction error, the parser may re-weight the validity of the structure built so far and rely less on structural cues during retrieval. Evidence from eye-tracking experiments in Parker and Phillips (2017) also supports their proposal that detection of agreement prediction error induces subsequent retrieval and interference. Specifically, the effect of grammaticality violations are reflected in early measures such as first-pass reading times, while the facilitatory interference effects are reflected in later measures.

Parker and Phillips incorporate this weighted cue-combinatorics scheme into their model for long-distance dependency processing based on the cue-based retrieval model of parsing under the Adaptive Character of Thought–Rational (ACT-R) architecture (Anderson, 1990; Lewis and Vasishth, 2005; Vasishth et al., 2008). Their model succeeds in simulating the facilitatory interference effects in ungrammatical English reflexive-antecedent dependencies. Thus, Parker and Phillips argue that cue weighting during retrieval depends on the predictability of linguistic dependencies. For predictable dependencies like subject-verb agreement, the neutralized structural cues allow facilitatory interference effects to surface in ungrammatical sentences. For unpredictable dependencies like reflexive-antecedent binding, the prioritized structural cues minimize facilitatory interference effects. We will call this model the "cue-preference-by-predictability" account for present purposes.

To summarize the theoretical landscape, in the face of empirical interference asymmetries, Tanner et al. (2014) propose that dependency predictability determines the need for retrieval (the retrieval-by-predictability account) while Parker and Phillips (2017) instead argue that dependency predictability affects cue-preference during retrieval (the cue-preference-by-predictability account). Ryu and Lewis (2021), on the other hand, suggest that word predictability based on noisy memory representations directly characterizes retrieval interference (the noisy-memory-based-predictability account; see also Futrell et al., 2020; Hahn et al., 2022).

Compared to previous emphasis on the predictability of a linguistic dependency during the retrieval operations (e.g., Dillon et al., 2013; Parker and Phillips, 2017; Tanner et al., 2014; Wagers et al., 2009), the predictability of the target lexical item itself has received less attention. However, target expectation is crucially implicated by the accounts reviewed above, where predictability modulates whether a target is retrieved, the cues prioritized in actualizing the retrieval, or the strength of the the memory representation being (re-)activated. In order to dissociate the effects of dependency predictability from that of target predictability, we manipulate target predictability in the present study. Specifically, we test whether the interference effects still occur under both grammatical and ungrammatical conditions for unpredictable dependencies such as noun-phrase ellipsis (as reported in Martin et al., 2012 and Tanner et al., 2014), when we additionally control for the predictability of the target item. We turn to this issue in the next section and identify specific EEG-based predictions concerning how target expectations modulate memory retrieval that tease apart these three hypotheses.

2.1.3 Interplay between retrieval interference and target expectations

In contrast to an extensive literature on cue-based retrieval interference, less work has focused on the interplay between similarity-based interference and probabilistic expectations of the target items. To our knowledge, Campanelli et al. (2018) presented the fist study to probe a related question. That is, they manipulated the probabilistic expectations of the word at the retrieval site, but not that of the target item itself. They used a dual-task design combining word list recall and sentence-reading to cross working memory load and main clause verb type, creating Baseline and Interference conditions. Working memory load was defined as whether or not participants needed to keep three nouns in memory while reading a stimulus sentence, and the main clause verb was varied as to whether or not the three nouns in memory could be a semantically compatible direct object for the main clause verb and thus cause interference.

Example (4) presents example Memory Load stimuli from Campanelli et al.; slashes mark segments of presentation during the self-pace reading task. (The No Memory Load conditions include the same sentences without the memory list.) Examples (b) and (c) illustrate Interference conditions because “website”, “handbag”, “password” in the memory list are plausible objects for the verb “create”, but not for “perform” as in the control condition illustrated in (a). The predictability of the main clause verb was varied by pairing it with different main clause subjects, thus creating High and Low Expectation conditions; (c) is the High Expectation condition since the verb “create” is highly expected in the context of the sentence with the subject “choreographer” while

(a) and (b) are Low Expectation conditions due to the lower expectation of the verbs "create" and "perform" with the subject "person".

(4) a. Low Expectation; No Interference
Memory list: **website-handbag-password**
It was the **dance**/that the person/who lived/in the city/**performed**/early last month.

b. Low Expectation; Interference
Memory list: **website-handbag-password**
It was the **dance**/that the person/who lived/in the city/**created**/early last month.

c. High Expectation; Interference
Memory list: **website-handbag-password**
It was the **dance**/that the choreographer/who lived/in the city/**created**/early last month.

Campanelli et al. report an expectation effect in reading times for both No Load and Load conditions at the spillover region (e.g., "early last month"). They observe faster reading time for the High Expectation (c) condition compared to Low Expectation ones (a,b). An interference effect in response time to comprehension questions was also observed for the Load conditions, with slower response time for the Interference conditions versus the No Interference one. This interference effect approached significance in reading time at the spillover region for the Low Expectation condition, but not for the High Expectation condition. Campanelli et al. argue that the higher expectation of words at the retrieval site might neutralize the interference effect on reading time. They conclude that sharp expectation offered by constraining context facilitates the retrieval and integration of previously encountered words at the retrieval site. The leading idea is that accumulated evidence may selectively pre-activate the target word and make it more available for retrieval compared to the distractors, consequently minimizing interference.

The observation of little interference effects under high expectation for grammatical sentences is also predicted by the retrieval-by-predictability account (Tanner et al., 2014), although their reason is the absence of retrieval, rather than facilitated retrieval suggested by Campanelli et al.. For the cue-preference-by-predictability account of Parker and Phillips (2017), in contrast, these grammatical contexts do not lead to any prediction error and thus should not affect how cues are weighted. Consequently, that account predicts there should be no modulation by expectedness.

Schoknecht et al. (2022) also examine the interplay between interference and expectation. They used German sentence pairs to compare two kinds of interference not yet discussed: retroactive interference where the distractor noun followed the target noun, and proactive interference where the distractor preceded the target nouns. As shown in example (5), in the first context sentence (a to d) of each sentence pair, the target noun is *Kfer* (“beetle” [masculine]), and the distractor noun *Wurm* (“worm”[masculine]) or *Raupe* (“caterpillar” [feminine]), with gender-matching distractor noun causing more interference. In the second target sentence (e) of each sentence pair, the target noun could be retrieved at the critical gender-marked article *den* (“the”[masculine]). The expectation level of this article was further differentiated by its cloze probability.

Schoknecht et al. find that retroactive interference elicits a broadly distributed negative ERP component compared to proactive interference for low, not high, expectation conditions. This negativity with fronto-central focus lasted from 300 to 500 ms after the article onset, and could indicate referential processing difficulty with multiple possible referent candidates (the “Nref” effect, discussed below in Section 2.1.5). Schoknecht et al. further argue that pre-activation of a fully predicted target word could prevent memory retrieval of the target and eliminate interference, consistent with the retrieval-by-predictability account (Tanner et al., 2014).

(5) a. High Retroactive Interference
In der Schachtel sitzt ein Kfer und im Glas liegt ein Wurm.
“In the box there sits a **beetle[masc.]** and in the glas there lies a **worm[masc.]**.”

b. High Proactive Interference
Im Glas liegt ein Wurm und in der Schachtel sitzt ein Kfer.
“In the glas there lies a **worm[masc.]** and in the box there sits a **beetle[masc.]**.”

c. Low Retroactive Interference
In der Schachtel sitzt ein Kfer und im Glas liegt eine Raupe.
“In the box there sits a **beetle[masc.]** and in the glas there lies a **caterpillar[fem.]**.”

d. Low Proactive Interference

Im Glas liegt eine Raupe und in der Schachtel sitzt ein Kfer.
"In the glas there lies a **caterpillar[fem.]** and in the box there sits a **beetle[masc.]**."

e. Target sentence for all conditions
Peter befreit den Kfer aus der Schachtel.
"Peter frees **the[masc.]** beetle[masc.] from the box."

These previous efforts help to clarify the interplay between expectation and memory retrieval, but face some limitations that we address in the present study. Note again that both Campanelli et al. (2018) and Schoknecht et al. (2022) manipulated the probabilistic expectations of the word at the retrieval site but not that of the target item. And while Campanelli et al. (2018) examined interference effects from a word list outside of the stimulus sentence under the dual-task paradigm, Schoknecht et al. (2022) focused on comparing proactive and retroactive interference, and did not directly assess the interaction between expectation and interference degree (high vs. low) which were strongly correlated in their materials. To better connect with previous studies on cue-based retrieval interference (e.g., Dillon et al. (2013); Jäger et al. (2020); Lago et al. (2015); Martin et al. (2012); Parker and Phillips (2017); Sturt (2003); Tanner et al. (2014); Wagers et al. (2009)), we directly manipulate the probabilistic exptectations of the target item and investigate interference effects from a distractor word within the stimulus sentence in a single-task sentence reading paradigm. We also modulate the degree of interference (high vs. low) separately from expectation.

2.1.4 Interference and expectations when resolving Mandarin ellipsis

Memory retrieval allows language users to establish a linguistic dependency between two non-adjacent constituents. This is especially crucial in the interpretation of so-called "silent" constituents as in linguistic ellipsis (Martin and McElree, 2009; Merchant et al., 2001). The ellipsis construction abounds in Mandarin (Li and Wei, 2014). To test how linguistic expectations modulate memory retrieval, we use Mandarin noun-phrase ellipsis construction shown in Table 2.1 in a single-task sentence reading paradigm.

Table 2.1: An example of Mandarin noun-phrase ellipsis.

媽媽	帶了	一件	**衣服**	在	**行李**	旁邊，	女兒	也	帶了	**一件**。
mma	dile	yjin	yf	zi	xngl	pngbin	nr	y	dile	yjin
mother	bring	one-CL$^{\text{Jian}}$	**shirt**$^{+Jian}_{+LocalObject}$	at	**luggage**$^{+Jian}_{-LocalObject}$	side	daughter	also	bring	**one-CL**$\{^{Jian}_{LocalObject}\}$
'The mother brought a **shirt** that was next to the **luggage**, and the daughter also brought **one**.'										

Mandarin is a numeral classifier language which means that a classifier is required between a noun and its preceding numeral, demonstrative, and certain quantifiers (Del Gobbo, 2014). The noun needs to agree with the classifier in semantic features. The classifier is a morpheme that "denotes some salient perceived or imputed characteristic of the entity to which an associated noun refers" (Allan, 1977); functionally, it serves to cognitively individualize and categorize units following semantic distinctions such as animacy, shape, orientation, rigidity, and nature/function (Croft, 1994). For example, the classifier 件 "**CL**$^{\textbf{Jian}}$" is used for individual objects such as shirts and luggage, while the classifier 本 "**CL**$^{\textbf{Ben}}$" is used for books and pamphlets.

In a sentential context, noun-phrase ellipsis can be licensed after the classifier (Cheng and Sybesma, 2014), as highlighted by the underlined number-classifier sequence in Table 2.1. Furthermore, because the noun-phrase ellipsis occurs in a coordinate sentence, the syntactic and semantic representations of the two conjuncts need to be symmetrical or parallel to each other according to the Parallelism Requirement (Zhang, 2009, p. 177). That is, 一件 "**one-CL**$^{\textbf{Jian}}$" from the sentence illustrated in Table 2.1 can only be interpreted as 一件衣服 "**one-CL**$^{\textbf{Jian}}$ **shirt**", and refers to a new referent under the category of **shirt**. In this way, 行李 "**luggage**" is a distractor noun; it is structurally illicit as the target antecedent because it does not reside in a symmetrical object position to the verb, but instead resides in a prepositional phrase adjunct. Successful licensing of noun-phrases ellipsis thus requires a structurally correct antecedent noun that also agrees with the classifier in semantic features.

When readers try to retrieve this target antecedent, the classifier critically provides structural cues and relevant semantic cues based on its agreement features. For notational purposes, the structural cue [**Local Object**] indicates a noun phrase that stands in the structural position of an object to a verb and does not reside within an adjunct; the semantic cue [**Jian**] stands for the semantic features of the classifier 件 "**CL**$^{\textbf{Jian}}$". In the present experiment, the target antecedent always matches the structural retrieval cue, while the distractor always mismatches the structural retrieval cue by virtue of its syntactic position. Crucially, both the target and distractor nouns may either match or mismatch the semantic cue of the classifier. Our use of a non-structural cue manipulation follows Martin et al. (2012) who manipulated determiner gender in Spanish noun-phrase ellipsis.[2]

To study interference effects during retrieval, our experimental design manipulates the semantic match between the classifier and a syntactically inaccessible distractor noun (see Table 2.3 below). We also vary the semantic match of the structurally correct target antecedent noun which offers a manipulation of sentence grammaticality. A mismatching target antecedent violates classifier-noun agreement and renders the sentence ungrammatical. Semantic cue-based interference from distractors is predicted by the cue-based retrieval theory (Lewis and Vasishth, 2005; Lewis et al., 2006; McElree, 2000; McElree et al., 2003; Van Dyke and Lewis, 2003), which emphasizes the effect of not only structural, but also non-structural cues when establishing linguistic dependencies like ellipsis. In grammatical sentences, cue-matching distractors could induce inhibitory interference and thus processing slow-downs (Franck et al., 2015; Jäger et al., 2017; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Van Dyke and McElree, 2011). (We discuss an alternative possibility further below that cue mismatch imposes processing difficulty, as suggested by Martin et al., 2012.) In ungrammatical sentences, facilitatory interference could be triggered by cue-matching distractors (Cunnings and Sturt, 2018; Dillon et al., 2013; Jäger et al., 2017, 2020; Lago et al., 2015; Parker and Phillips, 2017; Sturt, 2003; Tucker et al., 2015; Wagers et al., 2009).

Importantly, we test the influence of linguistic prediction on the retrieval operation by varying the lexical-semantic expectation of the target antecedent as determined by the main clause verb. The target word is less expected if the verb is congruent with both the distractor and target nouns. The target is highly expected if the verb is only congruent

[2] Other instances of semantic cues in the literature include animacy (Van Dyke, 2007; Van Dyke and McElree, 2011) and the semantic compatibility between verbs and their object nouns (Campanelli et al., 2018; Cunnings and Sturt, 2018; Van Dyke and McElree, 2006); these have been found to be diagnostic of similarity-based interference effects in both grammatical and ungrammatical contexts.

Table 2.2: Theoretical predictions from the retrieval-by-predictability and cue-preference-by-predictability accounts for interference effects under grammaticaliy and expectation manipulation for our experimental stimuli. Predictions from the noisy-memory-based-predictability account will be operationalized below in Section 2.3.2

		Retrieval-by-predictability account	Cue-preference-by-predictability account	Noisy-memory-based-predictability account
Grammatical	High Expectation	No interference	No interference	(operationalized in Section 4.2)
	Low Expectation	Inhibitory interference	No interference	
Ungrammatical	High Expectation	Facilitatory interference	Facilitatory interference	
	Low Expectation	Facilitatory interference	No interference	

with the target noun. The theories under consideration carry different predictions for how expectation should interact with interference and grammaticality in this construction. According to the retrieval-by-predictability account (Tanner et al., 2014), successful prediction minimizes both retrieval need and interference effects, as in grammatical noun-phrase ellipsis with highly expected target antecedent nouns. In contrast, less expected target nouns, or wrongly expected target nouns as in ungrammatical noun-phrase ellipsis, require the retrieval operation, during which the interference effects may surface. Those theoretical predictions are summarized in Table 2.2. Secondly, the cue-preference-by-predictability account (Parker and Phillips, 2017) suggests that when retrieval is a normal resolution process, prioritized structural cues could minimize the interference effects, as in grammatical noun-phrase ellipsis, or ungrammatical noun-phrase ellipsis with lowly expected target nouns. However, if retrieval is triggered by prediction error by highly expected target nouns in ungrammatical noun-phrase ellipsis, neutralized syntactic cues would increase the chance of interference effects (see Table 2.2 for summary). We have identified divergent predictions for our stimuli according to the retrieval-by-predictability (Tanner et al., 2014) and cue-preference-by-predictability (Parker and Phillips, 2017) accounts. Predictions of the noisy-memory-based-predictability account (Ryu and Lewis, 2021) for these stimuli are unknown. We operationalize that account below in Section 2.3.2 using the GPT2 large language model and ask whether surprisal values from GPT2 simulates responses reflecting interference and target expectations in a way similar to human data when reading the same stimulus items.

In sum, we manipulate expectation of the target antecedent by varying the main verb, grammaticality by manipulating semantic match of the classifier to target antecedent, and memory interference by manipulating the semantic feature of the distractor. This $2 \times 2 \times 2$ design yields two levels of expectation (High/Low), two levels of grammaticality (Grammatical/Ungrammatical) and two levels of interference (High/Low).[3] With this design, we first test the hypothesis that target expectations modulate interference effects under both successful and unsuccessful retrievals by establishing: (i) the ERP correlates of retrieval failure for ungrammatical ellipsis, (ii) ERP correlates of interference during retrieval success or failure, (iii) how contextual expectation might modulate these neural signatures. We then assess the explanatory power of three theoretical accounts: the retrieval-by-predictability account (Tanner et al., 2014), the cue-preference-by-predictability account (Parker and Phillips, 2017), and the noisy-memory-based-predictability account (Ryu and Lewis, 2021). To evaluate the noisy-memory-based-predictability account, we examine how well the GPT2 large language model captures the interference effects and the interaction between interference and target expectations, in parallel with the ERP analyses. The next section outlines our predictions according to previous literature on language-related ERP components that are relevant for cue-based retrieval.

2.1.5 The electrophysiology of retrieval mechanism and linguistic expectations

Research on similarity-based retrieval interference has primarily been built on behavioral data, including eye-tracking (Cunnings and Sturt, 2018; Dillon et al., 2013; Jäger et al., 2020; Parker and Phillips, 2017; Sturt, 2003; Van Dyke, 2007; Van Dyke and McElree, 2011), self-paced reading (Campanelli et al., 2018; Franck et al., 2015; Lago et al., 2015; Tucker et al., 2015; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Wagers et al., 2009), and speed-accuracy tradeoff (Van Dyke and McElree, 2011). The current study uses EEG for its high temporal resolution in detailing the retrieval process during sentence comprehension, whose multiple facets might not be fully captured by behavioral measures. As pointed out in Tanner et al. (2014), the lack of

[3] We label the conditions "High/Low Interference" by following e.g., Schoknecht et al., Van Dyke and Lewis (2003); Van Dyke (2007). And in line with the attraction literature, "High/Low Interference" corresponds to "Attractor-match/mismatch".

consistent behavioral interference effects reported in grammatical sentences might be due to its subtle nature, and EEG might offer measures sensitive enough to tease apart such nuances. Below we first review relevant ERP components before laying out predictions for our experimental design.

To begin with, semantically unexpected words induce a poststimulus negativity between 200 and 600 ms with a peak around 400 ms, which is largest over centro-parietal sensors and slightly right-lateralized (DeLong et al., 2005; Kutas and Hillyard, 1980, 1984; Kutas and Federmeier, 2011; Nieuwland et al., 2018, 2020). This "N400" effect has been linked to lexical-semantic processing difficulty (Kutas and Federmeier, 2011; Nieuwland et al., 2020) as expectation can facilitate semantic activation of a word due to pre-activation prior to the actual word appearance.

Secondly, violations of syntactic principles, including classifier-noun agreement violations (Hsu et al., 2014; Zhang et al., 2012), systematically induce a positive shift with a peak around 600 ms after stimulus onset and which is generally maximal over central posterior electrodes (Hagoort et al., 1993; Kaan et al., 2000; Kaan, 2002; Kaan and Swaab, 2003; Molinaro et al., 2011; Tanner et al., 2014; Yang et al., 2015). This "P600" effect has also been elicited by unfulfilled syntactic preferences in grammatically well-formed sentences (Friederici et al., 2001; Kaan and Swaab, 2003; Osterhout and Holcomb, 1992; Yang et al., 2010), and has been proposed to index revision processes prompted by inconsistency between an initial syntactic prediction and the received input. The latency of P600 may reflect the relative ease of the diagnosis, prior to actual reanalysis, for the revision operations (Friederici et al., 2001). The amplitude of the P600 increases with the degree of syntactic processing difficulty (Kaan et al., 2000; Kaan and Swaab, 2003). In addition, the P600 effect has been evoked by semantic verb-argument violations without syntactic violations or ambiguities (Kim and Osterhout, 2005; Kuperberg et al., 2003), and may indicate continuous combinatorial analysis in an effort to resolve conflicting semantic and syntactic representations (Kuperberg, 2007). Therefore, the onset and amplitude of the P600 can reliably signal syntactic processing difficulty, or combinatorial analytical effort, in both ungrammatical and grammatical sentences.

In investigating attraction effects in English subject-verb agreement, Tanner et al. (2014) reported a P600 effect to ungrammatical, relative to grammatical, verbs. They also reported a smaller P600 to ungrammatical verbs with a number-agreeing attractor noun, compared to those with a number-disagreeing attractor, consistent with a facilitatory interference effect. Tanner et al. interpreted the reduced P600 magnitude with a matching attractor to indicate less robust processing of the agreement violations. No significant difference was detected for grammatical conditions in that study. Similarly, Xiang et al. (2009) reported a P600 effect to ungrammatical, compared to grammatical, negative polarity item (NPIs) in English NPI licensing. A reduced P600 was also reported for ungrammatical NPIs with an intrusive licensor, relative to those without a licensor, showing a facilitatory interference effect. Relatedly, Martin (2018) used English verb-phrase ellipsis and found a P600 effect to ungrammatical, versus grammatical verbs, when the attractor verb-phrase mismatched the voice of the target verb-phrase. Martin suggested that the voice feature match between the attractor and retrieval cue in the ungrammatical condition was disruptive for antecedent retrieval and processing, but did not explicitly test for inhibitory or facilitatory interference effects. The P600 effect can thus effectively index both grammaticality and interference effects.

Martin et al. (2012), on the other hand, examined noun-phrase ellipsis marked by the gender-bearing determiner *otro/a* ("another") in Castilian Spanish. They found a sustained, broadly distributed negativity effect to ungrammatical, compared to grammatical, determiners. This resembles the "Nref" effect which is a sustained frontal negative deflection that onsets about 280 ms post-stimulus when an insufficiently specific referential expression, including nouns and pronouns, cannot identify a unique referent from multiple competitors according to previous discourse (e.g. Van Berkum et al., 1999; Van Berkum, 2009). Martin et al. further found this negativity to be elicited by grammatical determiners mismatching the gender of the attractors, in comparison to those matching the attractors. They argued that the attractors were temporarily considered as antecedent candidates, but did not fully match the retrieval cues. They suggest that the recency and/or similarity (to targets) of a "local agreement attractor" in memory affected retrieval success even in grammatical ellipsis with retrievable antecedents. Although not statistically significant, this negativity seemed to be smaller for ungrammatical determiners matching the gender of the attractors, compared to those mismatching the attractors, hinting perhaps at a facilitatory interference effect.

While limited, these findings provide some foundation for ERP predictions in the present study. We first predict a P600 effect to ungrammatical, relative to grammatical, ellipsis, due to syntactic violation of classifier-noun agreement. Alternatively, or additionally, ungrammatical (versus grammatical) ellipsis might drive an Nref effect as observed by Martin et al. in the case that no referent can be established for the elided noun.

Following Tanner et al. (2014) and Xiang et al. (2009), we expect a larger P600 (and/or Nref) effect in grammatical ellipsis for the cue-matching distractor, compared to the cue-mismatching one. This would be an instance of inhibitory interference as the semantic compatibility of the distractor might disrupt dependency formation for the target, and confuse the reader in settling on a unique antecedent and referent, when a search for antecedent is initiated at the numeral-classifier position. This inhibitory interference due to a cue-overload could lead to both syntactic and referential processing difficulty reflected in larger P600 and/or Nref. The account of Martin et al. (2012) carries a contrasting prediction for grammatical sentences: They hold that an effect in the opposite direction (i.e. larger P600 and/or Nref for *mis*-matching distractors) could follow if the ERP is driven by a mismatch between the distractor term and retrieval cues in memory. For ungrammatical ellipsis, the cue-matching distractor could reduce the P600 and/or Nref effect compared to the cue-mismatching one. This prediction follows under the theory that comprehenders might mistakenly consider the distractor as a valid target if it matches the semantic cue. Such an illusion of grammaticality (Wagers et al., 2009) could result in facilitatory interference, facilitating both syntactic and referential processing marked by smaller P600 and/or Nref.

Lastly, in grammatical ellipsis, an interference effect on the P600 and/or Nref could be neutralized when the target antecedent is highly predictable, as predicted by the retrieval-by-predictability account (Tanner et al., 2014). In ungrammatical ellipsis, the

P600 and/or Nref effect driven by facilitatory interference could be enlarged by a highly anticipated target antecedent. If a strong expectation of the target antecedent is defied at the retrieval site by a mismatching classifier, interference from the distractor would become more evident, in line with the cue-preference-by-predictability account (Parker and Phillips, 2017). In addition, the predictions of the noisy-memory-based-predictability account (Ryu and Lewis, 2021) on the experimental stimuli will be derived below, and compared with the human ERP data.

2.2 Methods

2.2.1 Materials

Forty four sets of eight conditions were created. To fulfill the expectation, grammaticality and interference manipulations, each condition in a set varied only in terms of: (1) the main verb that carried a high or low prediction for the antecedent, (2) the critical classifier (of the elided noun phrase) that semantically matched or mismatched the antecedent, and (3) the distractor that semantically matched or mismatched the critical classifier. See Table 2.3 for an example stimulus set and Section 2.8 regarding access to the complete set of experimental materials.

Using the classifier 件 "**CL$^{\textbf{Jian}}$**" at the retrieval site, (a) and (b) in Table 2.3 are Grammatical because the target nominal antecedent 衣服 "**shirt**" matches the semantic cue of the classifier, marked as [+**Jian**]. (a) is also High Interference since the intervening nominal distractor 行李 "**luggage**" also matches the semantic cue ([+**Jian**]), while (b) is Low Interference due to a mismatch ([−**Jian**]) by the distractor 書籍 "**book**". Due to the structural configuration, the target antecedent is always a match to the syntactic cue of the classifier ([+**Local Object**]), and the distractor a mismatch ([−**Local Object**]).

With the classifier 本 "**CL$^{\textbf{Ben}}$**" at the retrieval site, (c) and (d) are Ungrammatical since the target antecedent 衣服 "**shirt**" mismatches the semantic cue of the classifier ([−**Ben**]). (c) is also High Interference because the distractor 書籍 "**book**" is a match to the semantic cue ([+**Ben**]), and (d) is Low Interference as a result of a mismatch ([−**Ben**]) by the distractor 行李 "**luggage**".

Table 2.3: An example of experimental stimuli.

(a) Low Expectation; Grammatical; High Interference

媽媽	帶了	一件	**衣服**	在	**行李**	旁邊，	女兒	也	帶了	一件	出發	旅行。
mma	dile	yjin	yf	zi	xngl	pngbin	nr	y	dile	yjin	chf	lxng
mother	bring	one-CL$^{\text{Jian}}$	**shirt**$^{+Jian}_{+LocalObject}$	at	**luggage**$^{+Jian}_{-LocalObject}$	side	daughter	also	bring	**one-CL**$\{^{Jian}_{LocalObject}\}$	go.on	trip

‘The mother brought a **shirt** that was next to the **luggage**, and the daughter also brought **one** to go on a trip.’

(b) Low Expectation; Grammatical; Low Interference

媽媽	帶了	一件	**衣服**	在	**書籍**	旁邊，	女兒	也	帶了	一件	出發	旅行。
mma	dile	yjin	yf	zi	shj	pngbin	nr	y	dile	yjin	chf	lxng
mother	bring	one-CL$^{\text{Jian}}$	**shirt**$^{+Jian}_{+LocalObject}$	at	**book**$^{-Jian}_{-LocalObject}$	side	daughter	also	bring	**one-CL**$\{^{Jian}_{LocalObject}\}$	go.on	trip

‘The mother brought a **shirt** that was next to the **book**, and the daughter also brought **one** to go on a trip.’

(c) Low Expectation; Ungrammatical; High Interference

媽媽	帶了	一件	**衣服**	在	**書籍**	旁邊，	女兒	也	帶了	一本	出發	旅行。
mma	dile	yjin	yf	zi	shj	pngbin	nr	y	dile	ybn	chf	lxng
mother	bring	one-CL$^{\text{Jian}}$	**shirt**$^{-Ben}_{+LocalObject}$	at	**book**$^{+Ben}_{-LocalObject}$	side	daughter	also	bring	**one-CL**$\{^{Ben}_{LocalObject}\}$	go.on	trip

‘The mother brought a **shirt** that was next to the **book**, and the daughter also brought **one** to go on a trip.’

(d) Low Expectation; Ungrammatical; Low Interference

媽媽	帶了	一件	**衣服**	在	**行李**	旁邊，	女兒	也	帶了	一本	出發	旅行。
mma	dile	yjin	yf	zi	xngl	pngbin	nr	y	dile	ybn	chf	lxng
mother	bring	one-CL$^{\text{Jian}}$	**shirt**$^{-Ben}_{+LocalObject}$	at	**luggage**$^{-Ben}_{-LocalObject}$	side	daughter	also	bring	**one-CL**$\{^{Ben}_{LocalObject}\}$	go.on	trip

‘The mother brought a **shirt** that was next to the **luggage**, and the daughter also brought **one** to go on a trip.’

(e) High Expectation; Grammatical; High Interference

媽媽	穿了	一件	**衣服**	在	**行李**	旁邊，	女兒	也	穿了	一件	出發	旅行。
mma	chunle	yjin	yf	zi	xngl	pngbin	nr	y	chunle	yjin	chf	lxng
mother	wear	one-CL$^{\text{Jian}}$	**shirt**$^{+Jian}_{+LocalObject}$	at	**luggage**$^{+Jian}_{-LocalObject}$	side	daughter	also	wear	**one-CL**$\{^{Jian}_{LocalObject}\}$	go.on	trip

‘The mother wore a **shirt** that was next to the **luggage**, and the daughter also wore **one** to go on a trip.’

(f) High Expectation; Grammatical; Low Interference

媽媽	穿了	一件	**衣服**	在	**書籍**	旁邊，	女兒	也	穿了	一件	出發	旅行。
mma	chunle	yjin	yf	zi	shj	pngbin	nr	y	chunle	yjin	chf	lxng
mother	wear	one-CL$^{\text{Jian}}$	**shirt**$^{+Jian}_{+LocalObject}$	at	**book**$^{-Jian}_{-LocalObject}$	side	daughter	also	wear	**one-CL**$\{^{Jian}_{LocalObject}\}$	go.on	trip

‘The mother wore a **shirt** that was next to the **book**, and the daughter also wore **one** to go on a trip.’

(g) High Expectation; Ungrammatical; High Interference

媽媽	穿了	一件	**衣服**	在	**書籍**	旁邊，	女兒	也	穿了	一本	出發	旅行。
mma	chunle	yjin	yf	zi	shj	pngbin	nr	y	chunle	ybn	chf	lxng
mother	wear	one-CL$^{\text{Jian}}$	**shirt**$^{-Ben}_{+LocalObject}$	at	**book**$^{+Ben}_{-LocalObject}$	side	daughter	also	wear	**one-CL**$\{^{Ben}_{LocalObject}\}$	go.on	trip

‘The mother wore a **shirt** that was next to the **book**, and the daughter also wore **one** to go on a trip.’

(h) High Expectation; Ungrammatical; Low Interference

媽媽	穿了	一件	**衣服**	在	**行李**	旁邊，	女兒	也	穿了	一本	出發	旅行。
mma	chunle	yjin	yf	zi	xngl	pngbin	nr	y	chunle	ybn	chf	lxng
mother	wear	one-CL$^{\text{Jian}}$	**shirt**$^{-Ben}_{+LocalObject}$	at	**luggage**$^{-Ben}_{-LocalObject}$	side	daughter	also	wear	**one-CL**$\{^{Ben}_{LocalObject}\}$	go.on	trip

‘The mother wore a **shirt** that was next to the **luggage**, and the daughter also wore **one** to go on a trip.’

Employing the main clause verb 帶了 “**bring**”, (a) to (d) are Low Expectation because both the target antecedent 衣服 “**shirt**”, and the distractors 行李 “**luggage**” and 書籍 “**book**” can be a conceptually possible object noun of the verb. A change of the main clause verb to 穿了 “**wear**” makes (e) to (h) High Expectation since the verb is only semantically congruent with the target antecedent 衣服 “**shirt**”, but not with the distractors 行李 “**luggage**” or 書籍 “**book**”.

Target antecedent and distractor nouns were matched across conditions in terms of log frequency (base 10): antecedent noun $M = 3.75$ ($SD = 0.91$), different classification distractor 3.67 (0.66), same classification distractor 3.50 (0.80).[4] In addition, the numerals preceding the critical classifiers ranged from one to nine, with each number repeating for four or five times across sets. The critical classifier and its next word never occurred in a sentence-final position. Below we present the norming procedure for the expectation, grammaticality and interference manipulations based on the conditional probability of the object nouns given the verbs, and the classifier-noun collocation frequency.

To begin with, the conditional probability of the object nouns given the verbs was used to estimate whether the main verb highly predicted the antecedent noun compared to the distractor nouns (High Expectation condition), or whether the main verb did not predict the antecedent as well as the distractor nouns (Low Expectation condition). We used Chinese Word Sketch (Huang et al., 2005), a web-based system for collocation extraction based on the 1.4 billion words of Linguistic Data Consortium's Chinese Gigaword corpus (Huang, 2009).

For the High Expectation condition, the main verb always took as the object the antecedent noun more often than the distractor nouns, showing high expectation for the antecedent over the distractor, given the main verb. The raw counts that yield the conditional probability are: verb $M = 14877.11$ ($SD = 52665.99$), antecedent 462.68 (1829.16), different classification distractor 0.09 (0.0062), same classification distractor (raw count $= 0$). The conditional probability of the antecedent or distractor noun given the verb are: antecedent $M = 0.07$ ($SD = 0.12$), different classification distractor 0.000039 (0.00018), same classification distractor (raw count $= 0$).

Due to the size of the billion-word-corpus, our critical verbs and nouns are sparsely distributed in the data set. The key take-away is that the conditional probability of the antecedent given the high expectation verb is at least three orders higher than that of the distractor given the verb.

For the Low Expectation condition, the main verb showed no preference in taking the

[4] "Same classification distractors" refers to distractors that are semantically compatible with the same classifiers as the antecedents, and "different classification distractors" refers to those that are not.

antecedent or the distractors as object nouns. The raw counts that yield the conditional probability are: verb M = 71398.96 (SD = 106005.96), antecedent 62.32 (234.66), different classification distractor 11.77 (12.81), same classification distractor 26.25 (62.91). The conditional probability of the antecedent or of the distractor noun given the verb exhibited no significant differences: antecedent M = 0.0028 (SD = 0.0096), different classification distractor 0.0017 (0.0062), same classification distractor 0.0022 (0.0064).

The critical classifiers were selected from Gao and Malt (2009), who developed a taxonomy of 126 common individual classifiers and their associated object nouns. 41 different critical classifiers were used in the Grammatical condition, and 32 different critical classifiers in the Ungrammatical condition. 21 of them appeared in both conditions for different stimuli sets. The (in)congruity between the classifiers and object nouns, including the target antecedent and distractors, was confirmed by classifier-noun collocation frequency from Chinese Word Sketch (Huang et al., 2005).

Following Chan (2019), the degree of relative incongruity of the classifier-noun pair was calculated by the ratio of the mismatching classifier-noun collocation frequency to the matching classifier-noun collocation frequency. The lower the mismatching classifier-noun collocation frequency, the lower the incongruity ratio. An incongruity ratio below 25% was the threshold of incongruity for the mismatching classifier compared to the matching classifier.[5] For example, the collocation frequency for "CL$^{\text{Ben}}$-shirt" and "CL$^{\text{Jian}}$-shirt" were 0 and 463 respectively, resulting in an incongruity ratio of 0%, indicating that "CL$^{\text{Ben}}$" was the semantically mismatching classifier, and "CL$^{\text{Jian}}$" was the semantically matching classifier, for the noun "shirt". The incongruity ratio for antecedent was below 0.4% for all 44 stimulus sets; the incongruity ratio for different classification distractor was below 4% for all 44 sets; the incongruity ratio for same classification distractor was below 2% for 43 sets, and was 25% for 1 set.

All 44 stimuli sets met the above criteria. To balance the amount of available data against participant fatigue, three lists of 30 stimuli sets were generated from the original

[5] Chan used 10% as the incongruity ratio threshold. For our stimuli, all of the ratios were below 4% except for one, which was 25%, due to the smaller number of instances of that particular classifier-noun pair in the collocation corpus (Huang et al., 2005).

44 sets, so that 42 sets were used twice in different lists and two sets were used for all three lists. For each list, 30 stimuli sets of eight conditions were randomized with 104 filler sentences with various lengths and syntactic structures. The fillers were all grammatical and involved no semantic incongruity. Each participant saw one list of 344 sentences. We adopted this approach by following, for example, Bornkessel-Schlesewsky et al. (2011) and Schoknecht et al. (2022) in maximizing the amount of data which can be collected from each participant, compared to employing a Latin square design, given the considerable time and effort required for EEG recording preparation. To militate against order effects, we fully randomized each list differently for each participant.[6]

2.2.2 Metric for noisy-memory-based-predictability

Following Ryu and Lewis (2021), we used GPT2 surprisal as the metric to evaluate the noisy-memory-based-predictability account in relation to interference effects and the interaction between interference and target expectations. We computed the surprisal values of the critical number-classifier sequence from Chinese GPT2 (Du, 2019; Radford et al., 2019), a large-scale Transformer model pretrained on CLUECorpusSmall (Xu et al., 2020) with five billion Chinese words. Surprisal (Hale, 2001; Levy, 2008) was defined as the negative log probability of the critical number-classifier sequence given left-context of the stimulus sentence. We derived the surprisal values for all 44 stimuli sets.

2.2.3 Participants

Thirty native Mandarin speakers (6 Males, 24 Females; aged 19 to 37; mean age=23.7 years) participated in the study after signing a written consent form. The screening criteria include right-handedness, normal or corrected-to-normal vision, and no history of neurological or psychiatric disorders affecting the language functioning based on self report. All procedures were in accordance with protection for human subjects at the University of Michigan, following protocol HUM00081060. Participants were compensated 15 USD/hour.

[6] As the trial order effect might moderate the observed results, we conducted a post-hoc trial order Bayesian model analysis below in Section 2.3.1 to rule out potential confounds of repeated exposure.

2.2.4 EEG procedure

Participants were comfortably seated approximately 100 cm from the computer screen in an isolated room, and were instructed to silently read sentences while minimizing movement and eye blinks. Sentences were presented word-by-word at the center of the screen in white text on a black background using an RSVP paradigm with the stimuli segmented as shown in Table 2.3. A fixation cross of 500 ms began each trial and each word lasted for 300 ms with a 300 ms inter-stimulus-interval. One forth of the trials were followed by a yes/no comprehension question to check attentiveness, and the participants answered by pressing one of two keys on the keyboard with left or right index finger. The yes/no response keys were counterbalanced across participants. All participants showed above-chance accuracy on the comprehension questions (mean percent correct = 80%, range = [60%, 92%]). The next trial started immediately after the key press, or after 300 ms for trials without comprehension questions. After a short practice session to familiarize the participants with the task, the main session took approximately an hour, including 344 trials, with a break every 15 trials.

2.2.5 EEG recording

EEG was recorded from 32 actively-amplified electrodes, mounted on an elastic cap (actiCAP, EASYCAP GmbH), and arranged across the scalp following the Standard 32-channel actiCAP snap layout. Two electrodes were placed above and below the left eye to monitor vertical eye movements. The recordings were digitized at 500 Hz between 0.1 and 200 Hz, and referenced to the left mastoid electrode. Channel impedances were kept below 25 kOhms.

2.2.6 Data analysis

We used FieldTrip, a MATLAB toolbox (Oostenveld et al., 2011), for data processing. The EEG was high-pass filtered at 0.1 Hz, re-referenced to the average of the left and right mastoid electrodes, and epoched from 300 ms before the critical number-classifier sequence onset to 1000 ms after the onset. Trials with eye movements were removed using Independent Component Analysis (Jung et al., 2000), and trials with excessive noise were excluded based on visual inspection, with a removal range of 0.1%–19% (median = 10%) across participants. Data from channels exceeding the impedance threshold or introducing excessive noises were interpolated by surface spline interpolation (Perrin et al., 1987) (median channels interpolated per participant = 1, range = [0, 4]). Trials were low-pass filtered at 20 Hz, corrected with a 100 ms pre-stimulus baseline, and averaged together per condition.

In accordance with previous literature (Molinaro et al., 2011; Parker and Phillips, 2017; Sturt, 2003; Tanner et al., 2014; Xiang et al., 2009), we analyzed centro-posterior electrodes within 500–800 ms (for the main effect of grammaticality) and 600–800 ms (for interference effects) after the critical word onset (e.g., 一件 “**one-CL$^{\textbf{Jian}}$**”) for the predicted P600 effect. We followed Parker and Phillips (2017) and Sturt (2003) in dissociating the grammaticality and the interference effects, where the former temporally precedes the later. Specifically, while Tanner et al. (2014) reported grammaticality effect in the 500–800ms time interval, Xiang et al. (2009) reported grammaticality effect in the 400–600 and 600–800ms time intervals, and interference effect in the 600–800ms interval in centro-posterior electrodes. As proposed in the review paper of Molinaro et al. (2011), various subcomponents could interact in determining the P600 effects, evidenced by reports of positivities of differing time intervals. In addition, for the predicted Nref effect as reported in Martin et al. (2012), we followed Martin et al. in analyzing all EEG electrodes within 400–1000 ms after the critical word onset. Within these spatio-temporal windows, we employed a non-parametric statistical test (Maris and Oostenveld, 2007) to address the multiple comparisons problem which arises from sampling at multiple electrodes and time points. The cluster-based permutation tests were implemented as follows: (i) a repeated-measures t-test was conducted on each electrode-time pair. (ii) pairs with $p < 0.05$ were clustered based on temporal-spatial adjacency, and their t-values were summed. (iii) steps (i) and (ii) were repeated for 1,000 times by randomly assigning condition labels. (iv) clusters with $p < 0.05$ were considered as statistically significant. (v) clusters under High Expectation and Low Expectation conditions were compared respectively for

Grammatical and Ungrammatical conditions to determine interaction effects.

To further compare the performance of GPT2 with human EEG data, we also extracted single-trial EEG amplitudes, measured in microvolts, by averaging per trial from selected centro-posterior electrodes (Cz, CP1, CP2, Pz, P3, P4) within 600–800 ms after the critical word onset (e.g., 一件). We then conducted Bayesian statistical model analysis on the z-scores of single-trial EEG amplitudes, the expectation×grammaticalty×interference manipulation, and GPT2 surprisal using the `brms` package (Bürkner, 2017) and the `loo` package (Vehtari et al., 2017), with weakly informative priors to improve convergence and avoid overfitting (see Section 2.3.2).[7]

2.3 Results

2.3.1 EEG

The critical number-classifier sequences elicited a slow late positive shift in Ungrammatical conditions compared to Grammatical conditions. Figure 2.1 plots the grand average ERPs and 95% confidence intervals at electrode Pz for all eight conditions, grouped by Low Expectation (left panel), High Expectation (right panel), Low Interference (upper panel), and High Interference (lower panel). All Ungrammatical conditions show a late positive component with the exception of Ungrammatical; High Expectation; High Interference condition, which is the predicted condition to drive the facilitatory interference effect indexed by a reduced P600 amplitude. All corresponding topographic distributions are based on the 500 to 800 ms interval after the critical word onset. Figure 2.2 further plots the grand average ERPs and 95% confidence intervals collapsed across all Ungrammatical, or Grammatical conditions at electrodes Cz, Pz and Oz in the upper panel, and in the lower panel displays the corresponding topographic distributions in four adjacent

[7] Priors for regression coefficients were defined as $\mathcal{N}(0,1)$. In the Wilkinson-Rogers notation, the models were specified as:

1. `eeg ∼ expectation * grammaticality * interference + (1 + (expectation * grammaticality * interference) | participant),`
2. `eeg ∼ expectation * grammaticality * interference + gpt2_surprisal + (1 + (expectation * grammaticality * interference) + gpt2_surprisal | participant)`

time windows spanning from 100 to 1000 ms after the critical word onset. The positive going waveform begins around 400–500 ms after word onset and lasts to 700–800 ms, with maximal amplitude over centro-posterior electrode sites, and with little laterality preference. This main effect of grammaticality was statistically significant from 512–644 ms (cluster $p = 0.049$).

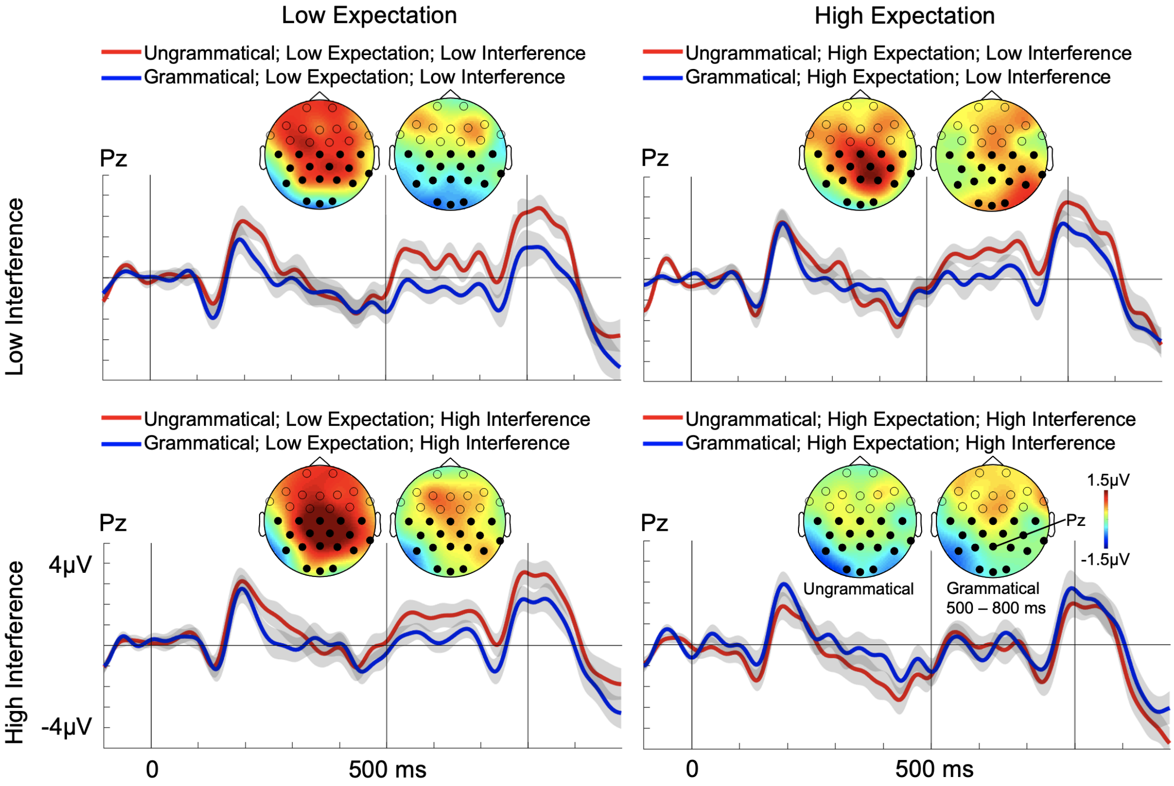

Figure 2.1: Grand averages and 95% confidence intervals (grey shading) elicited by the underlined critical word onset (i.e. number-classifier sequence) at Pz for all eight conditions grouped by Low Expectation (left), High Expectation (right), Low Interference (upper), and High Interference (lower). Ungrammatical condition is plotted with red lines and corresponds to the left scalp distribution for each contrast; Grammatical condition is plotted with blue lines and corresponds to the right scalp distribution. Scalp distributions are shown for the interval 500 to 800 ms after the critical word onset.

For Ungrammatical sentences, the High Interference condition elicited a reduced positivity relative to the Low Interference condition, but only when Expectation was High, not Low. Figure 2.3 shows the difference waves resulting from subtraction of the High Interference from the Low Interference condition and 95% confidence intervals, separately for High Expectation and Low Expectation conditions, along-side topographic difference maps for High Expectation and Low Expectation conditions respectively. The interac-

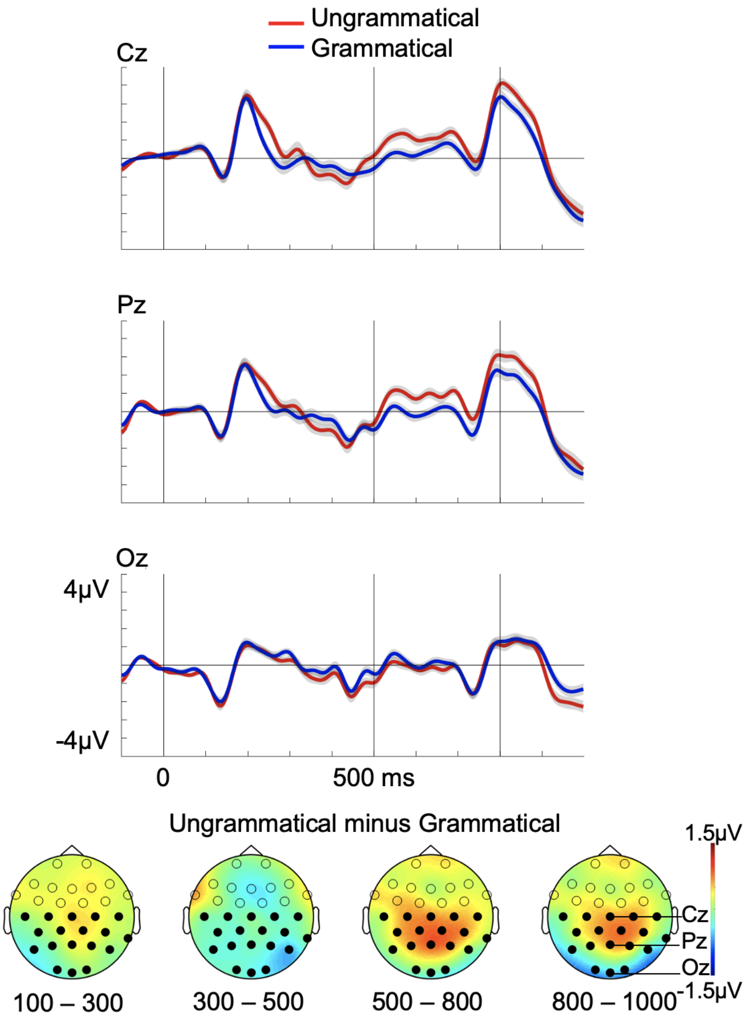

Figure 2.2: Grand averages and 95% confidence intervals (grey shading) at Cz, Pz, Oz for Ungrammatical (red lines) or Grammatical (blue lines) condition, and their scalp distributions during four consecutive time windows between 100 to 1000 ms after the underlined critical word onset (i.e. number-classifier sequence).

tion of Expectation, Grammaticality and Interference reached statistical significance from 678–748 ms ($p = 0.048$).

For Grammatical conditions, the High Interference condition elicited a more pronounced positivity compared to the Low Interference condition, but only in Low Expectation, not High Expectation conditions. Figure 2.4 shows difference waves resulting from subtraction of the High Interference from the Low Interference condition and 95% confidence intervals, for High Expectation and Low Expectation conditions respectively along with the topographic distributions for these differences waves separately for High Expectation and Low Expectation conditions. However, this trend did not reach statistical significance.

In contrast to the observed P600 effects, no Nref effect was detected when comparing Ungrammatical to Grammatical conditions ($p = 0.12$). And no modulation of Nref

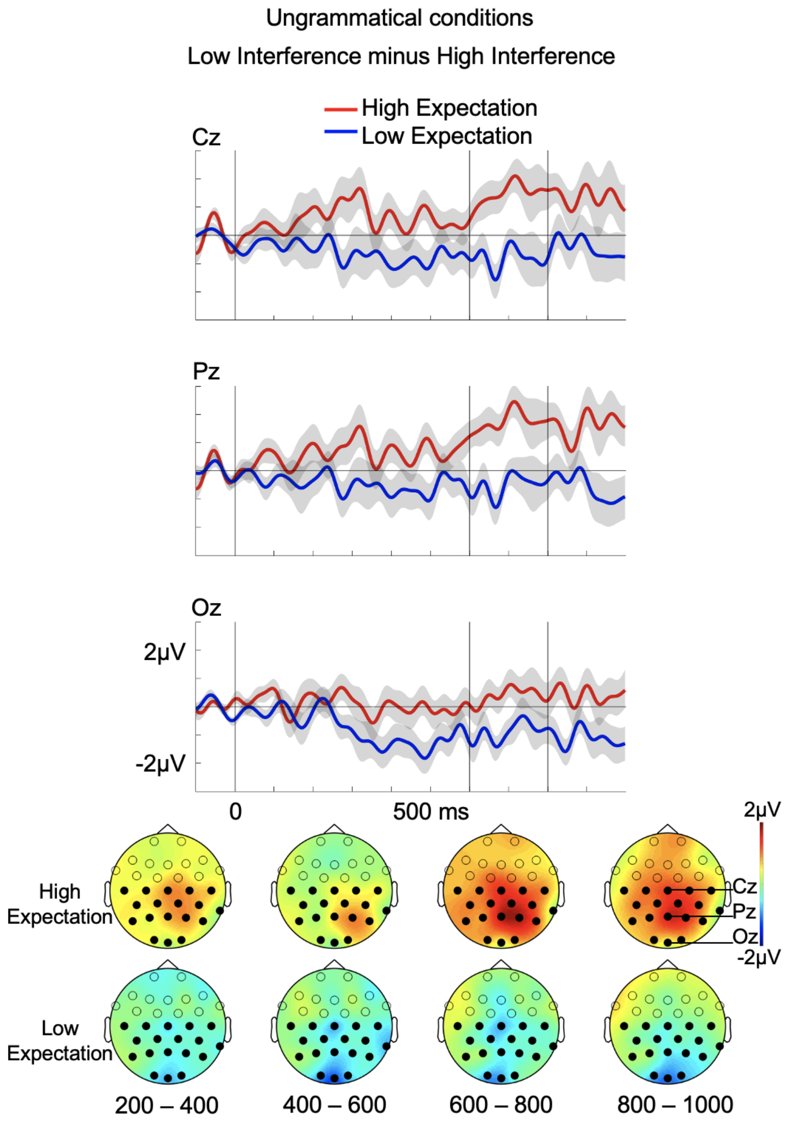

Figure 2.3: Difference waves and 95% confidence intervals (grey shading) at Cz, Pz, Oz for Low Interference minus High Interference in High Expectation (red lines) or Low Expectation (blue lines) ungrammatical condition, and their scalp distributions during four consecutive time windows between 200 to 1000 ms after the underlined critical word onset (i.e. number-classifier sequence).

effect by Interference and Expectation was found in either Ungrammatical ($p = 0.24$), or Grammatical ($p = 0.49$), conditions.

In brief, an ensemble of positive deflections consistent with the P600, but not Nref, was induced by Ungrammatical conditions, relative to Grammatical conditions. Moreover, in Ungrammatical conditions this P600 was reduced when the semantic features of an intervening distractor matched the classifier, but only when the target was highly predictable. This finding replicates and extends previous EEG findings on facilitatory interference (Martin et al., 2012; Tanner et al., 2014; Xiang et al., 2009), with a newly-established expectation modulation. For Grammatical conditions, a non-significant trend for prediction-modulated inhibitory interference is consistent with prior behavioral find-

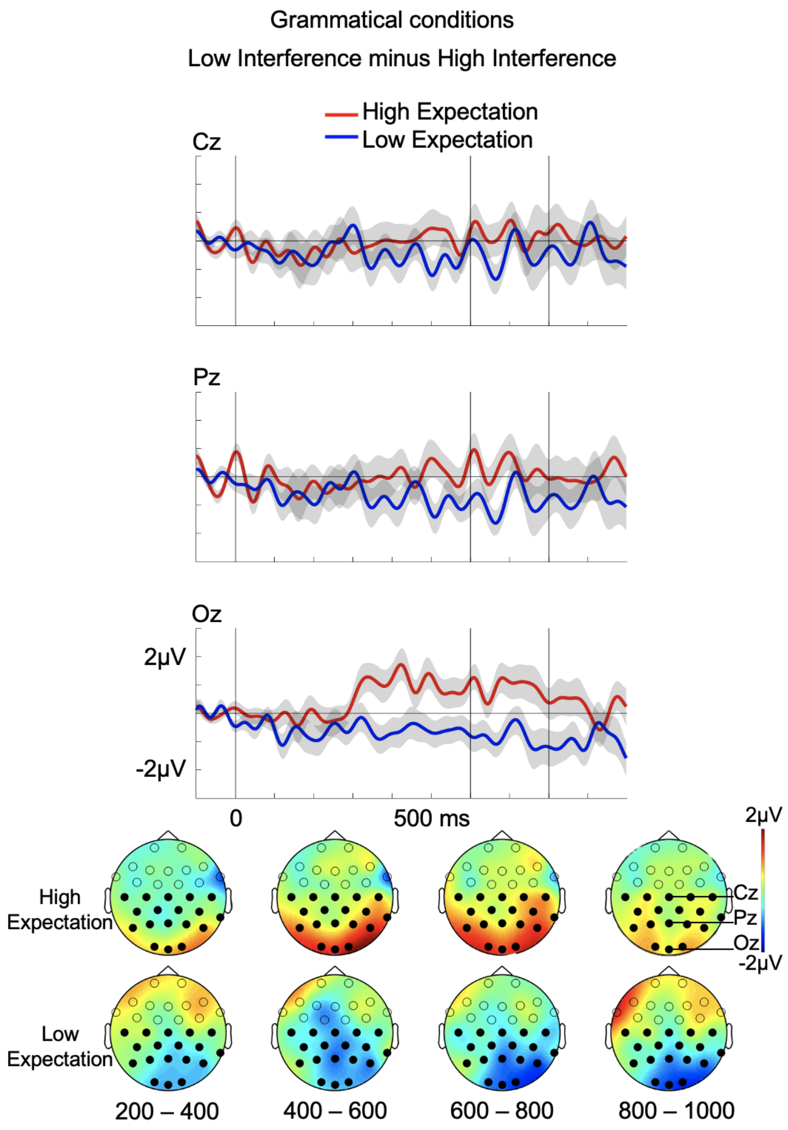

Figure 2.4: Difference waves and 95% confidence intervals (grey shading) at Cz, Pz, Oz for Low Interference minus High Interference in High Expectation (red lines) or Low Expectation (blue lines) grammatical condition, and their scalp distributions during four consecutive time windows between 200 to 1000 ms after the underlined critical word onset (i.e. number-classifier sequence).

ings (Campanelli et al., 2018).

Note that the choice of number-classifier sequences as the critical region presupposes that participants actively anticipate ellipsis; this assumption is confirmed by the grammticality effect observed at this region. Importantly for our purposes, the interference effect observed at this region is further modulated by expectation of the target antecedent.

We also analyzed the verb region immediately following the number-classifier sequence; it is this region where ellipsis is confirmed. We observed there a similar pattern of differences between conditions as in the critical region, although that pattern is indexed by a broadly distributed negativity rather than a P600. This negativity may be consistent with the Nref component. Specifically, a main effect of grammaticality was statistically significant from 400 to 828 ms ($p = 0.008$), with the Ungrammatical condition eliciting a broadly distributed negative shift relative to the Grammatical condition as plotted in Appendix A, Figure 1. For Ungrammatical conditions, an interference effect was statistically significant from 702 to 998 ms under High ($p = 0.028$), but not Low ($p = 0.41$) Expectation condition, with High Interference condition eliciting an enhanced negativity compared to the Low Interference condition (see Figure 2 in Appendix A). For Grammatical sentences, no interference effect was detected for either High ($p = 0.75$), or Low ($p = 0.34$) Expectation condition (see Figure 3 in Appendix A). The fact that interference effects did not surface in Low Expectation condition for both the critical number-classifier region, and also in the following verb region is consistent with the general finding of an expectation-modulated interference effect in ungrammatical sentences.

To evaluate the potential confound of repeated exposure, we also conducted a post-hoc Bayesian model analysis testing trial order effects at the critical number-classifier sequences. We fit a model with EEG single-trial P600 amplitudes as dependent variable and the trial order as independent variable along with the experimental factors and all higher order interactions. The 95% highest-posterior density (HPD) interval of the marginal effect of trial order spanned from -0.03 to 1.05 (mean = 0.54). Because this interval includes zero, we do not observe evidence suggesting trial order modulated the P600. Similarly, trial order did not reliably interact with the key experimental manipulation. We quantify this by examining the marginal HPD for the interaction between trial order and interference in each cell of the experimental design defined by gramamticality and expectation: Ungrammatical; High Expectation $M = -0.05$ $[-0.18, 0.09]$, Ungrammatical; Low Expectation -0.06 $[-0.20, 0.07]$, Grammatical; High Expectation 0.04 $[-0.08, 0.18]$, Grammatical; Low Expectation -0.06 $[-0.19, 0.07]$.

2.3.2 Noisy-memory-based-predictability

We now evaluate how well these EEG signatures of retrieval interference are captured by the GPT2 language model which we use to operationalize the noisy-memory-based-predictability account (Ryu and Lewis, 2021). Figures 2.5 and 2.6 plot averaged P600

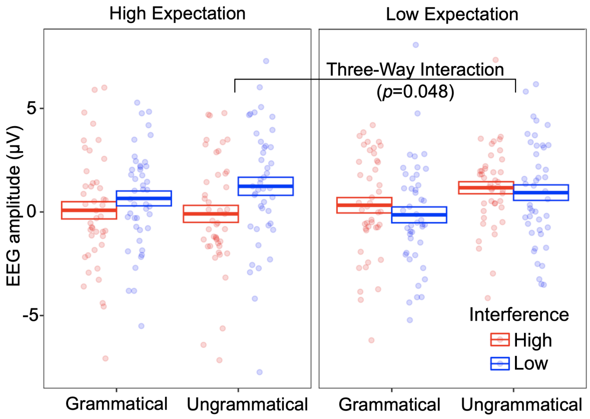

Figure 2.5: Stimulus-set-averaged EEG signal. EEG signals averaged per trial from selected centro-posterior electrodes (Cz, CP1, CP2, Pz, P3, P4) within 600–800 ms after the underlined critical word onset (i.e. number-classifier sequence) for High Interference (red points) or Low Interference (blue points) condition under High Expectation (left box) or Low Expectation (right box) condition for Grammatical (left within each box) or Ungrammatical (right within each box) condition. Points represent individual averages per stimulus set for plotting purposes, and boxes represent the group mean and 95% confidence intervals.

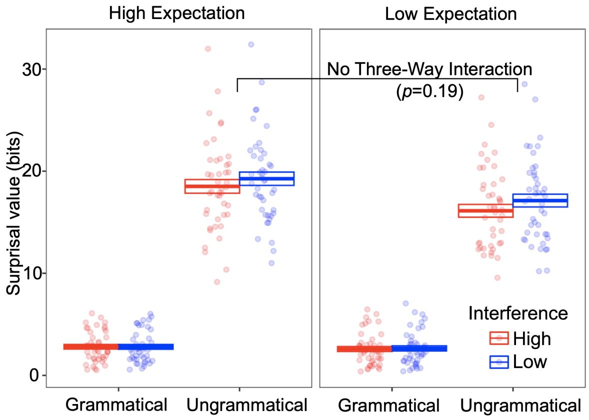

Figure 2.6: GPT2 surprisal. GPT2 surprisal at the critical number-classifier sequences across conditions as described in Figure 2.5. A three-way interaction is evident in EEG signals, but not GPT2 surprisal. This difference is statistically evaluated using model comparison in Section 2.3.2.

amplitudes from human EEG recordings along-side GPT2 surprisal values calculated at the critical number-classifier sequences. As detailed above, EEG results (Figure 2.5) exhibit a three-way interaction between grammaticality, interference and expectation. To repeat, for Ungrammatical conditions, the P600 was reduced in High (red points) vs. Low (blue points) Interference, but only in High (left box), not Low (right box), Expectation condition. No reliable difference surfaced in Grammatical conditions.

In contrast, GPT2 results (Figure 2.6) only show a two-way interaction between gram-

Table 2.4: ANOVAs for GPT2 surprisal at the critical number-classifier sequences.

Source	*dfs*	*F*	*p*
Expectation	1,43	11.67	<.01
Grammaticality	1,43	686	<.001
Interference	1,43	54.73	<.001
Expectation × Grammaticality	1,43	34.72	<.001
Expectation × Interference	1,43	15.27	<.001
Grammaticality × Interference	1,43	36.74	<.001
Expectation × Grammaticality × Interference	1,43	1.77	n.s.

Table 2.5: Theoretical predictions from retrieval-by-predictability account, cue-preference-by-predictability account and noisy-memory-based-predictability account for interference effects under grammaticaliy and expectation manipulation in our experimental stimuli.

		Retrieval-by-predictability account	Cue-preference-by-predictability account	Noisy-memory-based-predictability account
Grammatical	High Expectation	n.s.	n.s.	n.s.
	Low Expectation	Inhibitory interference	n.s.	n.s.
Ungrammatical	High Expectation	Facilitatory interference	Facilitatory interference	Facilitatory interference
	Low Expectation	Facilitatory interference	n.s.	Facilitatory interference

maticality and interference ($F(1, 43) = 36.74$, $p < .001$), without a three-way interaction with the additional expectation factor ($F(1, 43) = 1.77$, $p = 0.19$) (see Table 2.4). In particular, for Ungrammatical conditions, lower surprisal was observed in High (red points) vs. Low (blue points) Interference, under both High (left box) and Low (right box) Expectation conditions. No reliable differences occurred for Grammatical conditions. This two-way interaction replicates Ryu and Lewis, who simulated facilitatory, but not inhibitory, interference effects using GPT2 surprisal. The GPT2 results here serve to operationalize the theoretical predictions of the noisy-memory-based-predictability account for our experimental stimuli, which we summarize in Table 2.5, along with predictions from retrieval-by-predictability and cue-preference-by-predictability account (updating Table 2.2).

The absence of an interaction in GPT2 surprisal between Interference, Grammaticality and Expectation is bolstered by a single-trial Bayesian model analysis which found little contribution of GPT2 surprisal to the P600 response. Specifically, we compared two models as specified in Footnote 4. The first model has the z-scores of EEG single-trial signals as dependent variable, and the interaction among expectation, grammaticality and interference as independent variable, with random effect of the interaction among expectation, grammaticality and interference on each participant. The second model also has the z-scores of EEG single-trial signals as dependent variable, and the interaction among expectation, grammaticality and interference, as well as the z-scores of surprisal values, as independent variables, with random effects of the interaction among expectation, grammaticality and interference, and of surprisal on each participant. Models are compared via approximate leave-one-out cross-validation (Vehtari et al., 2017) and the comparison is summarized in terms of the difference in expected log pointwise predictive density ($\Delta ELPD$). The results showed that the first model without contribution of the GPT2 surprisal was the better performing model ($\Delta ELPD = -2.6, SE = 0.5$); the rubric of Sivula et al. (2020) describes values of ELPD < 4 as “small”.

In sum, while GPT2 surprisal does correlate with both grammaticality and facilitatory interference effects, it does not capture how expectations modulate these effects as found for the P600 ERP response. This is consistent with e.g. Hale et al. (2018), who reported no correlation between a surprisal measure derived from a recurrent neural network and the P600 (rather, surprisal was associated with an earlier anterior component in that study). Similarly, surprisal estimates derived from GPT2 are also limited in how well they predict human self-paced reading time and eye-gaze duration data compared to models which explicitly represent sentence structure (Oh et al., 2022) (see also Stanojević et al., 2023 regarding such limitations in capturing fMRI signals with a large language model.) Moreover, the larger the Transformer-based language model, the less predictive of human reading times the models become in terms of the surprisal estimates they generate (Oh and Schuler, 2023). This pattern follows in as much as the noisy-memory system that we operationalize with GPT2 might reflect memorization of sequences from immense amounts of training data (to maximize next-word prediction) in a way that is not human-like.

2.4 Discussion

We tested the hypothesis that target expectations modulate retrieval interference using Mandarin noun-phrase ellipsis construction in a single-task paradigm. We report three principle new findings: (i) ungrammatical noun-phrase ellipsis due to a mismatching classifier generates a P600 effect in comparison to grammatical ellipsis, (ii) the semantic feature of an intervening distractor, as well as the expectancy of the target antecedent, modulate the P600 signal in ungrammatical ellipsis, and (iii) a Transformer-based artificial neural network (GPT2) simulates the interaction between grammaticality and interference but does not capture the further modulatory effect of expectation. The predicted main effect of grammaticality in EEG signals indicates that grammatical and ungrammatical ellipsis are clearly differentiated by readers, and that the antecedent noun can be successfully retrieved and incorporated in grammatical conditions, but not in ungrammatical conditions. This effect also mitigates against the possibility that the local attractor is uniformly mistaken as the intended target for ellipsis resolution.

The observed interaction between grammaticality, interference and expectation in EEG signals highlights an interesting differential effect of cue-based retrieval interference and expectation on grammatical versus ungrammatical ellipsis. For ungrammatical ellipsis, the semantic cue-matching distractor in High Interference conditions might be temporarily taken as the antecedent noun and thus attenuate the P600 amplitude, compared to the Low Interference condition, but only when prediction error is incurred by a highly expected cue-mismatching target antecedent, not a less expected one. In contrast to this reliable interaction of expectation and facilitatory interference, the interaction of expectation and inhibitory interference effect in grammatical ellipsis is numerically smaller and not statistically reliable. In addition, GPT2 surprisal models only the facilitatory interference effects without the expectation modulation.

Observant readers might notice that the expectation manipulation in the current study varies not only the predictability of the target item, but also, inevitably, the semantic congruity of the distractor item with the main clause verb. That is, in High Expectation conditions, the target item is highly anticipated because the distractor items are semantically incongruent with the main clause verb. In Low Expectation conditions, the target item is less anticipated since the distractor items are also semantically compatible with the main clause verb. This design thus cannot disentangle the effects of target predictability from that of distractor congruity/plausibility. It is possible that both target predictability and distractor congruity/plausibility incur the prediction error in ungrammatical conditions, and subsequently moderate the facilitatory interference effects. While the discussion below focuses on the implication of target predictability, essential future work is needed to better differentiate the effects of target predictability from that of distractor congruity/plausibility. As discussed in Section 2.4.2 below, our ERP evidence (but not GPT2 simulation) is consistent with the cue-preference-by-predictability account (Parker and Phillips, 2017) developed under the broader cue-based retrieval framework. Below we first discuss our results in terms of the cognitive processes indexed by the observed ERP responses, and then relate them to theoretical models in terms of cue-based retrieval theory.

2.4.1 Neural signatures of grammaticality, interference and expectation

The mismatching classifiers in our study violate the agreement requirement of the antecedent nouns and evoke a P600 effect. Such an effect has been well-reported in response to violations of syntactic rules or expectation (Tanner et al., 2014; Xiang et al., 2009; Yang et al., 2015, 2010, inter alia), including classifier-noun agreement violation in Mandarin (Hsu et al., 2014; Zhang et al., 2012). Since the classifier mismatching the target antecedent disrupts ellipsis resolution, the parser may initiate a repair in syntactic relations, which is reflected in the P600 (Gouvea et al., 2010).

At the ellipsis site (i.e. the critical number-classifier sequence), we do not replicate the sustained negativity, or Nref effect, that was reported by Martin et al. (2012) for mismatching determiners in ungrammatical Spanish noun-phrase ellipsis. We speculate that the readers could readily detect the syntactic anomaly in the present study and engage in reanalyzing the syntactic relations, rather than directly experiencing referential failure when trying to establish a referent for the elided noun. At the verb region immediately following the number-classifier sequence, where ellipsis is confirmed, we do observe a broadly distributed negativity consistent with the Nref effect. We contend that this observation does not complicate the interpretation of the P600 effect observed in the previous region, and we do not speculate the underlying processes for this later negative component.

Importantly, the agreement violations indexed by P600 magnitude are attenuated by semantically matching distractor nouns in High Interference condition, relative to mismatching distractor nouns in Low Interference condition, when the highly expected target nouns do not match the classifiers. This attenuation is not present when nominal antecedents have low expectancy. Our results thus indicate that the resolution of noun-phrase ellipsis could be influenced when a local attractor matches the retrieval cue in an ungrammatical sentence where a highly expected antecedent noun could not be retrieved. The semantic feature of the local attractor and the expectancy of the target antecedent could both influence the retrieval process.

While Martin et al. reported a tentative (not statistically significant) facilitatory interference effect indexed by the Nref component during the resolution of noun-phrase ellipsis when a local attractor matches the gender of the ungrammatical determiner, Tanner et al. (2014) found a facilitatory interference effect indexed by a P600 when an attractor noun agrees in number with the ungrammatical verb during the formation of subject-verb agreement dependency. Similarly, Xiang et al. (2009) found a facilitatory interference effect indexed by a P600 during in a negative polarity construction (NPI) when an intrusive licensor intrudes for the ungrammatical NPI. We therefore first replicate the facilitatory interference effect reported in Tanner et al. (2014) and Xiang et al. (2009) for ERPs and extensively elsewhere for behavioral measures (see Introduction) and additionally demonstrate its interaction with expectation.

A numerically larger P600 also appears when both distractor nouns and less expected target nouns match the grammatical classifiers, compared to situation with mismatching distractors. No such increase surfaces in the presence of highly expected antecedents. This trend resembles the interaction between inhibitory interference effects and expectation indexed by reading time reported by Campanelli et al. (2018), who suggest smaller interference effect when target nouns are highly predicted and pre-activated by the preceding sentential context, thus facilitating target retrieval and minimizing distractor interference. Recognizing that this trend was not statistically reliable in that study, the pattern replicates the (marginally reliable) inhibitory interference effect and expectation interaction reported by Campanelli et al. with novel electrophysiological evidence.

Again, we do not replicate the Nref effect driven by mismatching distractors in grammatical noun-phrase ellipsis that was reported by Martin et al. (2012). As explained above, it is possible that distractor interference disrupts dependency formation for target antecedent, and first causes syntactic, not referential processing difficulty. We do, however, replicate the fact that interference effects can be detected, at least weakly, in both grammatical and ungrammatical noun-phrase ellipsis as in Martin et al.. Following Tanner et al. (2014), this interference effect might result from obligatory retrieval for noun-phrase ellipsis where agreement computation can not be predicted.

Taken together, the temporal resolution of ERPs enables investigation into the time course and underlying cognitive operations during the resolution of noun-phrase ellipsis. The P600 effect captures the interaction between grammaticality, interference and expectation, and indicates a modulation of syntactic processing difficulty by the three factors during sentence comprehension. In the next section, we further assess the retrieval-by-predictability account (Tanner et al., 2014), the cue-preference-by-predictability account (Parker and Phillips, 2017), and the noisy-memory-based-predictability account (Ryu and Lewis, 2021) according to the observed ERP results.

2.4.2 Theoretical implications for cue-based retrieval

Evaluated against the theoretical predictions summarized above in Table 2.5, our observed ERP interaction is most consistent with the cue-preference-by-predictability account (Parker and Phillips, 2017), but not the retrieval-by-predictability account (Tanner et al., 2014) or noisy-memory-based-predictability account as operationalized via GPT2 (Ryu and Lewis, 2021). According to the cue-preference-by-predictability account (Parker and Phillips, 2017), in ungrammatical noun-phrase ellipsis, the prediction error engendered by the highly predicted target might enable the distractor to interfere more easily and create an illusion of grammaticality, which then ameliorates the syntactic processing difficulty associated with the ungrammatical classifier. Since little prediction error would occur with the less predicted target, or in grammatical noun-phase ellipsis, interference from the distractor becomes less likely. We will review and extend the mechanism by which the cue-preference-by-predictability account of Parker and Phillips captures this interaction in the next section.

On the other hand, the retrieval-by-predictability account (e.g. Tanner et al., 2014) states that with the exception of highly predicted targets, less predicted or wrongly predicted targets need to be retrieved and that process is susceptible to distractor interference. This theory wrongly predicts interference effects with less expected targets, especially in ungrammatical contexts. Similarly, the noisy-memory-based-predictability account (Ryu and Lewis, 2021) as operationalized by the GPT2 language model fails to capture the observed EEG interaction by erroneously predicting interference effects with less predicted targets in ungrammatical sentences. We will briefly speculate the difference between the GPT2 language model and the extended cue-preference-by-predictability account in the next section, where we fit the results with one specific implementation of cue-based retrieval theory that incorporates information about predictability.

2.4.3 Predictability-dependent cue-weighting and cue-diagnosticity in retrieval mechanism

In this section we consider in more depth the theoretical implications of the interaction between grammaticality, interference and expectation we observe in human electrophysiological data. We first review below the cue-preference-by-predictability account (Parker and Phillips, 2017) developed under the cue-based retrieval framework, and subsequently extend their account to represent target expectation manipulated in our design. We will also return to the comparison of human EEG signals to neural-network based surprisal values, and end with a discussion of cue-diagnosticity.

Parker and Phillips proposed a retrieval mechanism with a cue-combinatorics scheme prioritizing structural cues over non-structural (e.g., morphological) cues, whose preferential weighting depends on predictability of the linguistic dependency. For unpredictable dependencies like reflexive-antecedent binding, retrieval is part of the normal resolution process, and syntactic cues are prioritized by default. In contrast, syntactic cues are not prioritized if retrieval is driven by a prediction error, and serves as a repair mechanism (Lago et al., 2015; Wagers et al., 2009), in predictable dependencies like subject-verb agreement. Due to the error signal, the parser might doubt the validity of the structures built so far, and thus minimize the use of syntactic cues for subsequent retrieval. This decreased importance of syntactic cues might in effect increase probability of interference from other cue-matching items that are structurally illicit for dependency resolution. This cue-preference-by-predictability account successfully predicts stronger facilitatory interference effects in subject-verb agreement compared to reflexive-antecedent dependency (i.e. type asymmetry) reported in the literature (Dillon et al., 2013; Jäger et al., 2017).

We propose that this cue-preference-by-predictability account (Parker and Phillips, 2017) can be naturally extended to capture the interaction between grammaticality, interference and expectation that we observe. Under their account, when the preceding main verb in our study highly predicts the target antecedent noun, but the mismatching classifier at the retrieval site violates the prediction, the parser would employ cue-based retrieval to find a matching noun to resolve ellipsis. The lower priority of structural cues during retrieval could therefore lead to facilitatory interference effect from the matching distractor noun. On the other hand, if the target antecedent remains less predicted, little prediction error would occur at the mismatching classifier, and structural cues would still be weighted more strongly during the retrieval, reducing the likelihood of local attraction. Taken together, we extend the theory of Parker and Phillips by showing that the cue-weighting scheme could additionally be affected by predictability of the to-be-retrieved linguistic item. Prediction errors generated at the retrieval site could neutralize the structural cues even in unpredictable dependencies such as noun-phrase ellipsis.

In support of the current view, the temporal profile of our EEG effects effectively replicates previous reading time experiments (Lago et al., 2015; Parker and Phillips, 2017). Specifically, we observed that the P600 effect of grammaticality violations (512–644 ms) numerically preceded that of facilitatory interference (678–748 ms), suggesting an initial detection of an agreement prediction error, followed by error-prone retrieval of the antecedent. (Note, though, that we did not statistically evaluate latency differences in this study.) This consistency, more broadly, points in favor of extending this cue-weighting scheme of Parker and Phillips to similarity-based interference effects in not only English and Spanish (Lago et al., 2015; Parker and Phillips, 2017), but also Mandarin.

Additionally, we extend Ryu and Lewis (2021) and demonstrate that Transformer-based neural network language models also predict certain pattern of facilitatory interference effect during Mandarin ellipsis processing. However, GPT2 appears to be insensitive to the modulatory effect of antecedent predictability on interference, which is key to human electrophysiological data explained by cue-based retrieval theory with preferential cue-weighting scheme (Parker and Phillips, 2017). As detailed above, the parser prefers syntactic cues over non-structural cues; thus facilitatory interference tends to emerge more when structural cues are neutralized by prediction error caused by a highly expected antecedent mismatching the retrieval cues. The fact that GPT2 predicts facilitatory interference effect across-the-board suggests a lack of such cue preference.

Finally, we will briefly discuss how predictability-dependent cue-weighting relates to cue-diagnosticity during the retrieval process. Since the classifier provides both structural and semantic cues at the retrieval site, our results implicate that successful retrieval of the cognitive target antecedent depends on whether retrieval cues could fully and unambiguously map onto the target antecedent (Van Dyke and Lewis, 2003). The strength of cue-diagnosticity for the target item could be weakened by other cue-matching items. According to Nairne (2002), the probability of retrieval of a target item depends on the extent to which retrieval cues match the target item relative to other possible candidates in memory.

Following the variant of cue-based retrieval model of Parker and Phillips (2017), we suggest that retrieval cues may be differentially weighted according to the expectancy of the target lexical item as manipulated in the current study. In particular, structural cues are weighted more than non-structural cues by default. Since only the targets, not the distractors, match the structural cue, this preferential cue-weighting strengthens cue-diagnosticity for the targets. However, if prediction errors are incurred by a highly predicted target, the preferential weighting no longer exists, thus weakening the cue-diagnosticity for the targets. As a consequence, the probability of target retrieval will be decreased while increasing the probability of facilitatory interference from distractor(s), as in the High Expectation/Ungrammatical/High Interference condition in the current study.

Preferential cue-weighting directly impacts cue-diagnosticity for ungrammatical ellipsis due to the presence of prediction error, but this account does not readily extend to grammatical ellipsis. This is because under that account, cue-reweighting is only motivated by ungrammaticality. Our results indicated a trend, albeit not statistically reliable, for predictability effect on grammatical inhibitory interference. If such a trend warrants theoretical analysis, we suspect that it might follow from pre-activation of target items (as suggested in e.g., Campanelli et al., 2018; Schoknecht et al., 2022) which increases their base-level activation (see Lewis and Vasishth, 2005; Parker and Phillips, 2017; Vasishth et al., 2008). The probability of target retrieval will thus be increased while the probability of inhibitory interference from distractor(s) is decreased, as observed here for High Expectation/Grammatical/High Interference sentences.

Our findings, broadly, provide evidence that structurally illicit local attractors as well as antecedent expectancy could affect noun-phrase ellipsis licensing and referential resolution. Successful retrieval to compute ellipsis during language comprehension is a function of the feature match between retrieval cues and target lexical item, relative to the feature match between retrieval cues and recent lexical item, modulated by the expectancy of the target.

2.5 Conclusion

Language users often need to access past memory representations during online sentence processing. Our EEG study demonstrated that the retrieval of a cognitive noun-phrase antecedent to license ellipsis could be affected not only by an intervening noun-phrase distractor, but also by the expectancy of the antecedent, especially when the sentence was ungrammatical. Our results align with the cue-preference-by-predictability account (Parker and Phillips, 2017), and support the claims that grammatical and semantic constraints can be implemented as retrieval cues, and that linguistic dependencies are resolved using a cue-based retrieval mechanism operating on content-addressable memory representations (Lewis and Vasishth, 2005; Lewis et al., 2006; McElree, 2000; McElree et al., 2003; Van Dyke and Lewis, 2003). The GPT2 transformer-based neural network language model does not show evidence of a similar cue-weighting mechanism. Building on previous English and Spanish studies, our Mandarin data also provide evidence that similar memory retrieval mechanism could be employed across languages in comprehension (Lago et al., 2015).

2.6 CRediT authorship contribution statement

Tzu-Yun Tung: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing —original draft, Writing —review editing. **Jonathan R. Brennan:** Formal analysis, Funding acquisition, Methodology, Resources, Software, Visualization, Supervision, Writing —review editing.

2.7 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

2.8 Data availability statement

The experimental stimuli, ERP data, and surprisal data from GPT2 are available online at `https://osf.io/xf3sy/?view_only=be336cd2b6e048b9b6aa56b07d4cc29f`.

2.9 Acknowledgments

We would like to thank the editor and the reviewers for their constructive feedback, which greatly improved the manuscript. We also appreciate the stimulating discussions with the audience of The 35th Annual Conference on Human Sentence Processing (HSP2022), The 14th Annual Meeting of the Society for the Neurobiology of Language (SNL2022), and the psycholinguistics research group at the University of Michigan.

2.10 Funding

This work was supported in part by Rackham Humanities Research Fellowship and Rackham Graduate Student Research Grants of the University of Michigan to Tzu-Yun Tung, and by the National Science Foundation Grant#1607251 to Jonathan R. Brennan.

2.11 Appendix A. Supplementary material

Supplementary data, including additional result plots, the experimental stimuli, ERP data, and surprisal data from GPT, to this article can be found online at https://osf.io/xf3sy/?view_only=be336cd2b6e048b9b6aa56b07d4cc29f.

CHAPTER III

Memory Retrieval and Predictions during Naturalistic Dependency Resolution

3.1 Introduction

Language comprehension is affected by the success of memory retrieval of previously encountered lexical items and also by the predictability of the word being processed given the prior context. The retrieval of a previous lexical item is theorized to depend on the time elapsed since that item first appears (i.e., the decay effect), and also on the occurrence of other similar items in working memory causing interference (i.e., the similarity-based interference effect; see Chapter II) (Arnett and Wagers, 2017; Cunnings and Sturt, 2018; Dillon et al., 2013; Franck et al., 2015; Glaser et al., 2013; Jäger et al., 2017, 2020; Lago et al., 2015; Martin et al., 2012; Martin, 2018; Mertzen et al., 2023; Sturt, 2003; Tucker et al., 2015; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2007; Van Dyke and McElree, 2011; Vasishth et al., 2019; Wagers et al., 2009; Xiang et al., 2009). On the other hand, contextual information routinely shapes how well a word can be anticipated and integrated, as discussed in Chapter I (Boston et al., 2008; Brothers and Kuperberg, 2021; Chen and Hale, 2021; Chow et al., 2018; DeLong et al., 2005, 2014; Demberg and Keller, 2008; Federmeier and Kutas, 1999; Federmeier et al., 2007; Frank et al., 2015; Hale, 2001; Henderson et al., 2016; Kamide, 2008; Kutas and Hillyard, 1984; Levy, 2008; Levy and Keller, 2013; Roark et al., 2009; Van Berkum et al., 2005; Wicha et al., 2004; Xiang and Kuperberg, 2015).

While memory retrieval and predictability have been extensively investigated in the sentence processing literature, this work joins a small, but growing set of studies examining these two facets of processing together (Campanelli et al., 2018; Futrell et al., 2020; Parker and Phillips, 2017; Ryu and Lewis, 2021; Schoknecht et al., 2022; Tanner et al., 2014). Importantly, previous work, including the experiment reported in Chapter II, uses mainly artificially-constructed individual sentences as experimental stimulus, which diverge from the language use in everyday life. It remains to be seen whether the effects of memory retrieval and predictions also surface during a more naturalistic language setting such as audiobook listening (Brennan et al., 2016; Brennan, 2016; Brennan and Hale, 2019; Hale et al., 2018; Lopopolo et al., 2017; Willems et al., 2016).

The sorts of sentences that demand retrieval, and present possible instances of decay and interference, are common in every-day language. For instance, example (1) presents the direct English translation of an excerpt from the audiobook story "Le Petit Prince". Here, the nominal subject **Little prince** is encoded in memory and needs to be retrieved upon reaching the main clause verb **<u>asked</u>** to interpret "little prince", but not "me", as the agent of the action "asked".

(1) **Little prince** to **me** **<u>asked</u>** many questions.

The longer the linear distance between the verb and the subject noun, the less activated in working memory the subject noun becomes when reaching the verb site. The decay of the subject noun could thus hinder its retrieval success. Moreover, the retrieval site provides the syntactic cue "grammatical subject" that maps onto the memory representation of "little prince", but not that of "me", as the structurally licensed target subject (e.g., see Arnett and Wagers, 2017; Dillon et al., 2013; Kush and Phillips, 2014; Kush et al., 2015, 2017, for using structural information as a constraint to retrieve target noun-phrases from memory). The semantic cue "animate", on the other hand, matches both nouns, which could cause interference during retrieval. Interference could result in misretrieval of the semantic-cue-matching distractor "me", and misinterpretation of "me" as the agent of the action "asked". In addition, the relative likelihood of the critical verb **<u>asked</u>** relies on the conditioning sentential context. A high-constraint context can preactivate the word and facilitate its access from long-term memory during processing.

The current study will focus on establishing whether the comprehension of subject-verb dependency during human natural listening can be characterized by: (i) the cue-based working-memory retrieval theory (Lewis and Vasishth, 2005; Vasishth et al., 2008) implemented under the Adaptive Control of Thought–Rational (ACT-R) architecture (Anderson, 1990), and/or (ii) the word predictability derived from noisy memory representations such as those in a large language model (Futrell et al., 2020; Ryu and Lewis, 2021).

3.1.1 Cue-based working-memory retrieval

Cue-based working-memory retrieval has been proposed to mediate the resolution of long-distance dependency, leading to processing complexity (Lewis and Vasishth, 2005; McElree, 2000; Vasishth et al., 2008), as demonstrated in the work presented in Chapter II with Mandarin noun-phrase ellipsis (Tung and Brennan, 2023). This retrieval mechanism is constrained by independently motivated principles of memory and cognitive skills, and specifically applied to human sentence processing. Under this theory, sentence parsing consists of a sequence of memory retrievals, which are affected by both fluctuating activation and similarity-based interference. Processing difficulty arises when the retrieval target has low activation level in memory due to longer time elapsed since it last appeared, or when the target cannot be easily distinguishable from other similar items in memory. These cognitive principles have been formulated computationally under the Adaptive Control of Thought–Rational (ACT-R) architecture (Anderson, 1990), which serves as a formal model of word-by-word sentence comprehension.

However, previous experiments use mostly highly-controlled discrete sentences to test the theory, and less is known about the generalizability of the framework to naturalistic language. This study thus aims to test the hypothesis that cue-based retrieval as characterized by ACT-R also underlies language processing in a more naturalistic setting such as audiobook listening. To apply this framework to naturalistic text, we will first parse and annotate our audiobook stimulus with the Stanford Neural Network Dependency Parser (Chen and Manning, 2014). Within the stimulus, we focus on subject-verb dependency because it has previously been heavily studied with controlled experiments (see Jäger et al., 2017, for review) but not in a more naturalistic setting. After the parser identifies all subject-verb dependencies in the stimulus, the ACT-R metric predictions will be obtained for all target verbs in the dependencies. ACT-R can be divided into different components which provide quantitative estimates of factors that affect activation, such as the amount of interference. In addition, there are several hyper-parameters that are subject to debate in the literature, such as the relative weighting of different cues (see Section 2.1.2 in Chapter II for more discussion). I will thus be evaluating a suite of models, with different parameter settings corresponding to different theoretical proposals, and also for each model I will derive a set of metrics that captures sub-parts of the retrieval process. Those models and metrics will be statistically assessed with the EEG data recorded at the target verb region. We will further compare those models with the model based on word predictability (to be described below), and rank all models according to their statistical fit against the EEG data. The ranking will determine whether the ACT-R models or the predictability model could better predict human electrophysiological signals during subject-verb dependency resolution.

While traditional cue-based models of parsing employ structural and non-structural retrieval cues simultaneously and equally to solve resolve dependencies by default (Lewis and Vasishth, 2005; McElree, 2000; Vasishth et al., 2008), recent work has considered the possibility that cues may be weighted differently or may be evaluated at different time lags. Structural cues are either weighted more strongly than (Parker and Phillips, 2017; Van Dyke and McElree, 2011; Yadav et al., 2022), or evaluated prior to (Mertzen et al., 2023; Sturt, 2003), non-structural cues. In other words, syntactic constraints could serve as an "early but defeasible filter" when accessing the retrieval targets (Sturt, 2003). These proposals thus predict weaker interference effects from distractor items mismatching the syntactic cues, at least in an early processing stage. The current study will compare ACT-R variants with equal and differential cue-weighting scheme in terms of their predictive power for human electrophysiological signals.

Cue preference could depend on a variety of variables. Firstly, the cue-preference-by-predictability account (Parker and Phillips, 2017; Tung and Brennan, 2023, also see P. 13 in Chapter II) highlights the predictability of linguistic dependencies and retrieval targets. According to this account, syntactic cues are prioritized by default during the retrieval operation if no prediction error occurs to neutralize this priority. This could be true for all grammatical contexts (without any grammatical violation) as well as for ungrammatical contexts with unpredictable target nouns or unpredictable dependencies such as reflexive-antecedent binding (as little prediction leads to error signals). The priority of the syntactic cues maximizes the success in retrieving the structurally relevant target items. In contrast, prediction error could be incurred by predictable target nouns or predictable dependencies such as subject-verb agreement in ungrammatical situations. The parser could consequently reevaluate the validity of the structure under construction, and no longer prioritize the structural cue, giving structurally inaccessible distractor items a chance to exert interference effects.

The cue-preference-by-predictability account accurately characterizes the empirical findings of consistent interference effects in ungrammatical sentences with subject-verb agreement, and the occasional interference effects in ungrammatical sentences with reflexive-antecedent dependency (Dillon et al., 2013; Jäger et al., 2017). In their eye-tracking experiments on ungrammatical reflexive licensing, Parker and Phillips (2017) reported interference effects only when the target word had 2 morphological-cue-mismatches, but not when the target had 1 mismatch. This pattern was later successfully simulated with the syntactic cues being weighted 1.6 times higher than the morphological cues. The syntactic cues did not serve the gating function (cf. Van Dyke and McElree (2011)) as their preferential weighting still allowed structurally inaccessible distractor items to be considered for retrieval when the target word mismatched more morphological cues.

Secondly, cue preference could relate to individual variations. Yadav et al. (2022) reported that only fast readers weighted structural cues higher. Cue weighting could thus be related to reading proficiency and language experience. We will call this the "cue-preference-by-proficiency account" to highlight its particular features. Finally, cue preference could be attributed to cross-linguistic variations. When contrasting German with English, Mertzen et al. (2023) found that semantic cues seemed to enter the computation slightly slower than syntactic cues, and associated the superiority of syntactic cues over semantic ones to the richer morphosyntactic marking in German. We name this the "cue-preference-by-morphosyntax account" to contrast with other accounts.

The above accounts can be operationalized differently for our experiment. First of all, with the assumption that little prediction error would be generated by the grammatical audiobook texts, structural cues would by default receive stronger weighting during the retrieval process. The cue-preference-by-predictability account (Parker and Phillips, 2017; Tung and Brennan, 2023) thus predicts that the ACT-R model with preferential cue-weighting would outperform that with equal cue-weighting in simulating our experimental results. In Section 3.2.2 below, we will construct an ACT-R model with preferential cue-weighting (ACT-R-2), which I contrast with an ACT-R model with equal cue weighting (ACT-R-1). Secondly, with the working hypothesis that all native speakers have high native language proficiency, the participants would prefer structural over non-structural cues when processing their native language. The cue-preference-by-proficiency account (Yadav et al., 2022) similarly predicts that the model with preferential cue-weighting would be the better performing model. Thirdly, assuming that Mandarin resembles more with English in terms of their less explicit morphosyntactic marking, structural cues would not be more superior than non-structural cues for the Mandarin retrieval operation. Consequently, the cue-preference-by-morphosyntax account (Mertzen et al., 2023) predicts that the model with equal cue-weighting would capture our human data better than the model with preferential cue-weighting.

Table 3.1 summarizes the above accounts with their proponents and operationalization for our experiment. Importantly, recent development has seen an integration of predictability into the memory-based ACT-R models, which is evident in the cue-preference-by-predictability account (Parker and Phillips, 2017; Tung and Brennan, 2023). Section 3.1.2 below will further discuss the incorporation of memory constraints into predictive models for next-word prediction, and operationalize the noisy-memory-based-predictability account (Ryu and Lewis, 2021; Tung and Brennan, 2023) for our experiment, to be contrasted with the above theories.

Under the cue-based memory retrieval framework, in order to resolve the long-distance subject-verb agreement dependency, the subject noun needs to be retrieved upon encountering the main verb (Dillon et al., 2013; Glaser et al., 2013; Mertzen et al., 2023; Parker and Phillips, 2017; Van Dyke and Lewis, 2003; Van Dyke, 2007; Van Dyke and McElree, 2011). Aside from the requirement of positional configuration, the licensed subject noun also agrees with the dependent main verb in terms of semantic-pragmatic properties (e.g., animacy) (Mertzen et al., 2023; Van Dyke, 2007; Van Dyke and McElree, 2011). The verb thus provide both syntactic cues and semantic cues in search of a compatible subject noun.[1] If other structurally illicit distractor nouns also match those semantic retrieval cues, retrieval interference may occur. Such semantic interference effect has been reported using both offline and online processing measures, including paraphrasing (Stolz, 1967), comprehension accuracy (King and Just, 1991), self-paced reading (Van Dyke, 2007), speed-accuracy tradeoff (Van Dyke and McElree, 2011), and eye-tracking (Mertzen et al., 2023; Van Dyke, 2007; Van Dyke and McElree, 2011). While prior literature mainly focuses on English and German, the current study aims to achieve cross-linguistic validation and assess whether the retrieval of a subject noun-phrase will be impacted by an animacy-matching distractor noun during the resolution of subject-verb dependency in Mandarin.

3.1.2 Interaction between memory retrieval and predictability

Along-side working memory considerations, probabilistic expectations about sentences have also been proposed to account for processing difficulty in real-time comprehension of natural language. Surprisal (Hale, 2001; Levy, 2008), based on conditional probability, is one incremental information-theoretic complexity metric that quantifies

[1] Such semantic cues have also been used for the verbs and their associated object nouns in the cleft or object relative clause construction to investigate the interference phenomenon (Campanelli et al., 2018; Cunnings and Sturt, 2018; Van Dyke and McElree, 2006).

processing complexity of a sentence in terms of word-by-word expectations. While "lexical surprisal" calculates how surprising a word is given the context, "syntactic surprisal" assesses how surprising the syntactic structure to be generated is, in order to integrate the current word into the sentence (Roark et al., 2009).

Surprisal has been shown to successfully predict human word-by-word comprehension in diverse linguistic tasks using different methodological measures. The tasks included comprehension of naturalistic paragraphs from novels or newspaper (Brennan et al., 2016; Brennan and Hale, 2019; Demberg and Keller, 2008; Hale et al., 2018; Henderson et al., 2016; Lopopolo et al., 2017; Willems et al., 2016), of single sentences from novels presented in random order (Frank et al., 2015), of constructed narratives with syntactically complex sentences (e.g., sentential embeddings, relative clauses and non-local dependencies) (Roark et al., 2009), and of constructed individual sentences (Boston et al., 2008; Hale, 2001). Those studies employed techniques such as electroencephalography (Brennan and Hale, 2019; Frank et al., 2015; Hale et al., 2018), fMRI (Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; Willems et al., 2016), eye-tracking (Boston et al., 2008; Demberg and Keller, 2008) and self-paced reading (Roark et al., 2009). Surprisal can therefore estimate the degree to which individual words are pre-activated in the brains of language users.

Rather than being viewed as two independent processing mechanisms, recent efforts have explored how both memory retrieval and prediction may interact in conditioning language comprehension efforts. For example, both unidirectional transformer-based language models such as Generative Pre-trained Transformer-2 (GPT2) (Radford et al., 2019), and variants such as the lossy-context surprisal model of Futrell et al. (2020) can be used to compute the surprisal value of a lexical item based on the noisy memory representations of its preceding context. With some information affected by noise, the noisy memory representation provides an incomplete version of the true context. The information-loss is a general feature of memory representations, and may result from cognitive resource constraints, as argued in the resource-rational model of fine-grained memory representations (Hahn et al., 2022). Surprisal from GPT2 has been

Table 3.1: Summary of the cue-preference-by-predictability, cue-preference-by-proficiency, cue-preference-by-morphosyntax and noisy-memory-based-predictability accounts with their proponents and operationalizations for our experimental stimuli.

Account	Cue-preference-by-predictability	Cue-preference-by-proficiency	Cue-preference-by-morphosyntax	Noisy-memory-based-predictability
Proponent	Parker and Phillips (2017) Tung and Brennan (2023)	Yadav et al. (2022)	Mertzen et al. (2023)	Ryu and Lewis (2021)
Operationalization	ACT-R with preferential cue-weighting (ACT-R-2)	ACT-R with preferential cue-weighting (ACT-R-2)	ACT-R with equal cue-weighting (ACT-R-1)	Word predictability formalized as surprisal

further proposed to characterize the retrieval interference effects in English sentences with subject-verb agreement and reflexive-antecedent dependencies (the "noisy-memory-based-predictability account" (Ryu and Lewis, 2021; Tung and Brennan, 2023)). For our experiment, the account thus predicts that word predictability derived from the artificial neural network model with noisy memory representations could characterize the human processing profile. Table 3.1 presents the proponent and operationalization of the noisy-memory-based-predictability account, along with that from the cue-preference-by-predictability, cue-preference-by-proficiency and cue-preference-by-morphosyntax accounts introduced in Section 3.1.1.

To tease apart these accounts, the current study will use the ACT-R and word-predictability metrics to model the incremental parsing operations separately, and to test whether any of those metrics could better detect the neural correlates of the parsing operations in an EEG experiment. The analysis below will focus on a particular aspect of the EEG signal: the sustained anterior effects with a post-stimulus onset of 280 ms. This choice reflects prior work on memory effects in language with EEG. That prior work (Martin et al., 2012) indicates that memory retrieval modulates a particular EEG response called the "Sustained Anterior Negativity". This effect is similar to the "Nref" effect (Van Berkum, 2009), elicited when a unique referent cannot be identified from multiple candidates in memory based on prior discourse.

3.2 Methods

3.2.1 Materials

We adopted materials used in "Le Petit Prince multilingual naturalistic fMRI corpus" (Li et al., 2022) to enable direct comparison between the neuroimaging and electrophysiological data. The stimuli were the first section of the Chinese "The Little Prince" audiobook (`http://www.xiaowangzi.org/`), which amounted to nine minutes, and was read by a professional female Chinese broadcaster. This section contained 2,427 words in 134 sentences with an average length of 18.11 words per sentence ($SD = 13.67$). Following Li et al., the auditory narrative was accompanied by visual drawings as appeared in the original text to aid in comprehension. Three drawings was respectively shown at the 10, 35 and 60 second timepoints of the section, and lasted for 15, 20 and 15 seconds. This choice of naturally occurring and contextualized, rather than artificially crafted and isolated, stimuli enhances the generalizability of the results.

3.2.2 Three models of memory retrieval and predictability

We use the ACT-R model of sentence processing (Lewis and Vasishth, 2005; Vasishth et al., 2008) to estimate memory cost in this more natural stimulus. To further probe the modulatory effect of cue-combinatorics on memory retrieval, we build two ACT-R variants with equal and differential cue-weighting scheme. These models serve to tease apart the cue-preference-by-predictability (Parker and Phillips, 2017; Tung and Brennan, 2023), cue-preference-by-proficiency (Yadav et al., 2022) and cue-preference-by-morphosyntax (Mertzen et al., 2023) accounts. The third model formalizes the expectation of a word via surprisal (Hale, 2001; Levy, 2008) using Chinese GPT2 (Du, 2019; Radford et al., 2019), and tests for the noisy-memory-based-predictability account (Ryu and Lewis, 2021; Tung and Brennan, 2023). We describe the mathematical formulation of each model below.

The original ACT-R model is based on Equation 3.1, which calculates the overall activation level A of the target lexical item i, A_i. A_i affects the probability of the item's retrieval and its retrieval latency.

$$A_i = B_i + \sum_{j=1}^{m} W_j S_{ji} \tag{3.1}$$

The base-level activation of the lexical item, B_i, in Equation 3.1 is specified in Equation 3.2. t_j stands for the time since j th retrieval of the item, while the decay parameter d is conventionally set to be 0.5 as default. Adding up the time for all n retrievals to the power of the negative decay parameter, and transforming it through natural logarithm, produce B_i. As a result, the more frequent an item occurs, or is retrieved, the higher its base-level activation becomes. This is called "activation boost".

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{3.2}$$

For the second term in Equation 3.1, the weight W for each retrieval cue j, W_j, equals G/j. G represents all available goal activation and defaults to 1 in ACT-R. The associative strength between a retrieval cue and the lexical item S_{ji} is further computed by Equation 3.3. The constant S is set to be 1.5 following previous modeling work on the fan effect (Anderson, 1990; Vasishth et al., 2008). The term fan_j expresses the number of lexical item matching the retrieval cue j. Consequently, the more lexical items matching the retrieval cue, the weaker the associative strength between the cue and the target item, which gives rise to retrieval interference.

$$S_{ji} = S - \ln\left(fan_j\right) \tag{3.3}$$

We further vary the weight for each retrieval cue W_j to create two ACT-R models. ACT-R-1 has the default equal cue-weighting for structural and non-structural cues. ACT-R-2 has the preferential cue-weighting for structural cues, which is 1.6 times higher than the non-structural cues, following Parker and Phillips (2017).

Lastly, we calculate the surprisal value of the critical word at the retrieval site (i.e., the verb in subject-verb dependency) using Equation 3.4.

$$surprisal(w) = -log_2(p(w|c)) \tag{3.4}$$

Surprisal (Hale, 2001; Levy, 2008), measured in bits, is the negative logarithm (base 2) probability p of a word w given a linguistic context c. Words with lower conditional probability convey more information and demand more cognitive load in sentence processing, resulting in higher surprisal values. The linguistic context includes all preceding words in the same sentence as the critical verb at the retrieval site. To enable direct comparison with prior work (Ryu and Lewis, 2021), we use Chinese GPT2 (Du, 2019; Radford et al., 2019) to derive the "lexical surprisal", that is, the conditional probability of a word based on its lexical identity (cf. syntactic surprisal).

The accuracy of each of these models in representing human psychological processes is the main focus of this paper. We will test the model predictions against human electrophysiological signals recorded during audiobook listening.

3.2.3 Stimulus annotations

The quantitative predictions of the ACT-R and surprisal metrics were tested against Mandarin human EEG data to investigate the effects of memory and predictability during the resolution of subject-verb dependency in language processing. In the following steps, to acquire the quantitative predictions for all stimuli, the first section of the Chinese "The Little Prince" audiobook was parsed by the Stanford Neural Network Dependency Parser (Chen and Manning, 2014) to identify all subject-verb dependencies and their intervening distractor nouns. For the ACT-R metric, the time-interval between subject and verb was noted to model activation decay. The structural feature [**Local Subject**] was annotated for the target subject; the semantic feature [**Animacy**] of the target subject and distractor nouns was annotated to query retrieval interference effects between nouns due to a shared animacy status (both are [+**Animate**] or [−**Animate**]), following previous literature (e.g., Mertzen et al., 2023; Van Dyke, 2007; Van Dyke and McElree, 2011). Interference was higher where the animacy feature matched between target and distractor nouns. Brought together, the activation level of the target subject noun-phrases was calculated by Equation 3.1, which served as a metric of memory cost. In addition, to obtain the expectation-based metric, the surprisal value of the critical verb

Table 3.2: Token counts of the: (1) High Interference, (2) Low Interference, and (3) No Interference configurations with animate or inanimate target subject noun-phrases in Mandarin subject-verb dependency.

Configuration		Count
(1) High Interference	Animate Target	40
	Inanimate Target	1
(2) Low Interference	Animate Target	31
	Inanimate Target	1
(3) No Interference	Animate Target	122
	Inanimate Target	28
Total		223

at the retrieval site was computed by Equation 3.4. These metrics were then statistically assessed against the human EEG data.

To illustrate the different ways retrieval may play out in this stimulus, consider the following three configurations for subject-verb agreement that we observed in this story: (1) High Interference configuration with at least one matching distractor, (2) Low Interference configuration with at least one mismatching distractor and no matching distractor, and (3) No Interference configuration without any distractor. Those three discrete configurations serve the illustration purpose of the retrieval interference effects only, while the statistical analysis uses continuous variables based on numerical ACT-R and word-predictability metrics defined in Section 3.2.2. Table 3.2 presents the token counts of each configuration with animate or inanimate target subject noun-phrases. 41 tokens of High Interference configuration, 32 of Low Interference, and 150 of No Interference amounts to 223 tokens in total.

We will provide concrete example sentences for each configuration, along with illustrations of the ACT-R metric derivation for each sentence below. To begin with, Figure 3.1 displays sample parse and annotations of three example sentences for the three key configurations. In the first clause of the first sentence, the subject **little prince** ([+**Animate**]) needs to be retrieved upon reaching the verb **ask** to establish the dependent relation between the nominal subject and the verb. Here, a distractor noun **me** ([+**Animate**]) intervenes between the subject and the verb. Due to a shared animacy status with the target subject ([+**Animate**]), the distractor has a high probability of causing similarity-based inhibitory interference effect in this grammatical sentence, hence the (1) High Interference configuration.

In the second sentence, the retrieval of the subject noun **you** ([+**Animate**]) occurs at the verb **fall**. Crucially, the intervening distractor **sky** ([−**Animate**]) has a low probability of generating inhibitory interference effect since it has a different animacy status compared to the target subject. This is the (2) Low Interference configuration. Finally, in the last clause of the first sentence, where the subject **he** ([+**Animate**]) depends on the verb **have**, no nominal distractor intervenes, which exemplifies the (3) No Interference configuration.

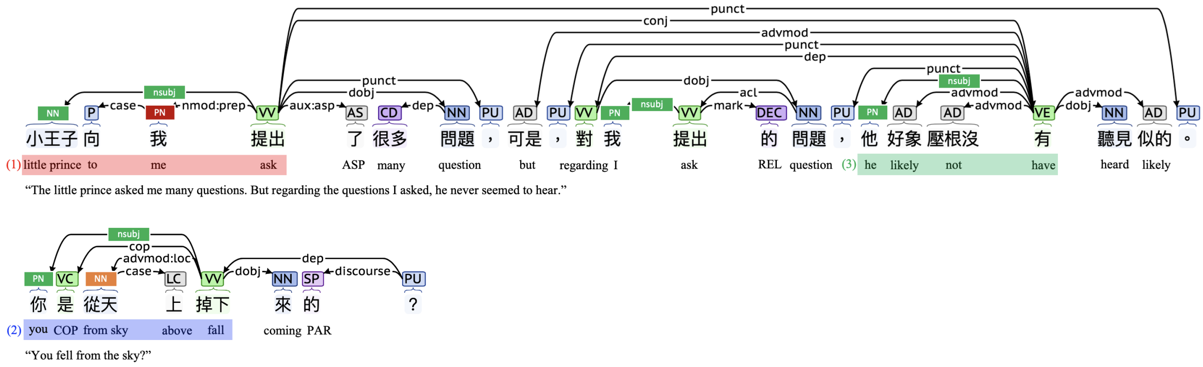
punct
conj
advmod
punct
dep
punct
dobj
nsubj
nsubj
case
nmod:prep
aux:asp
dep
dobj
acl
mark
nsubj
advmod
advmod
advmod
dobj
NN P PN VV AS CD NN PU AD PU VV PN VV DEC NN PU PN AD AD VE NN AD PU
小王子 向 我 提出 了 很多 問題 ， 可是 ， 對 我 提出 的 問題 ， 他 好象 壓根沒 有 聽見 似的 。
(1) little prince to me ask ASP many question but regarding I ask REL question (3) he likely not have heard likely
"The little prince asked me many questions. But regarding the questions I asked, he never seemed to hear."
nsubj
cop
advmod:loc
case
dobj
dep
discourse
PN VC NN LC VV NN SP PU
你 是 從天 上 掉下 來 的 ?
(2) you COP from sky above fall coming PAR
"You fell from the sky?"

Figure 3.1: Sample parse and annotations of the: (1) High Interference (red shading), (2) Low Interference (blue shading), and (3) No Interference (green shading) configurations in Mandarin subject-verb dependency.

Following the parse and annotations, the activation level of the target subjects in the three example sentences can be computed by Equation 3.1. For the (1) High Interference configuration example, we calculate the overall activation level A_i of the subject noun "little prince" at its dependent verb "ask". To get the the base-level activation B_i, we first subtract the onset timepoint of "little prince" (418.51 second) from that of "ask" (419.51 second) to estimate the time t_j since last retrieval of "he" (1 second).[2] Entering this time t_j into Equation 3.2 yields the base-level activation B_i (0). Since both the subject noun "little prince" and the intervening distractor "me" match the semantic retrieval cue [**Animacy**] between the target and retrieval site in the sentence, fan_j equals 2. This

[2] In the current model, we only assess the time t_j since the latest retrieval of the target subject, given the scope of the sentence the subject noun occurs in. The dependent relationship between a subject noun and the verb is defined by the Stanford Neural Network Dependency Parser (Chen and Manning, 2014).

obtains the associative strength S_{ji} between the retrieval cue and the target subject of 0.81 according to Equation 3.3. The subject uniquely matches the structural retrieval cue [**Local Subject**], leading to fan_j of 1 and associative strength S_{ji} of 1.5. S_{ji} can be multiplied with a weighting W_j of 1 for equal cue-weighting scheme, or with W_j of 1.6 for preferential cue-weighting. For ACT-R-1 with equal cue-weighting, overall activation level A_i (2.31) results from the the summation of the base-level activation B_i (0) and the sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ for the two retrieval cues (2.31). For ACT-R-2 with preferential cue-weighting, overall activation level A_i (3.21) is the summation of the base-level activation B_i (0) and the sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ for the two retrieval cues (3.21). Those values are summarized in Table 3.3.

Similarly, for the (2) Low Interference configuration example, we subtract the onset timepoint of the subject noun "you" (450.48 second) from that of "fall" (451.27 second) to get the time t_j since last retrieval of "you" (0.79 second), as well as the base-level activation B_i (0.12). With only the subject "you" matching the semantic cue [**Animacy**], the fan_j equals 1, and the associative strength S_{ji} 1.5. And with only the subject "you" matching the structural cue [**Local Subject**], the fan_j equals 1, and the associative strength S_{ji} 1.5. Combining B_i (0.12) with sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ (3) results in the the overall activation level A_i (3.12) for ACT-R-1. ACT-R-2 has the overall activation level A_i of 4.02 from the summation of B_i of 0.12 and sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ 0f 3.9 (see Table 3.3).

Finally, for the above example of (3) No Interference configuration, we subtract the onset timepoint of the subject noun "he" (422.92 second) from that of "have" (423.58 second) to measure the time t_j since last retrieval of "he" (0.66 second), which results in the base-level activation B_i of 0.21. The fan_j for the semantic retrieval cue [**Animacy**] is 1 because only the subject noun "he" matches the cue, and the associative strength S_{ji} 1.5. The fan_j for the structural cue [**Local Subject**] is 1 because of a uniquely matching subject, and the associative strength S_{ji} 1.5. For ACT-R-1, the overall activation level A_i is 3.21 by adding up B_i (0.21) and sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$

Table 3.3: The base-level activation B_i, sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$, and overall activation level A_i of the target subjects for ACT-R-1 and ACT-R-2, as well as the surprisal value of the critical verbs, in the three example sentences for the three key configurations.

Configuration	Target Subject	Critical Verb	ACT-R-1			ACT-R-2			Surprisal
			B_i	$\sum_{j=1}^{m} W_j S_{ji}$	A_i	B_i	$\sum_{j=1}^{m} W_j S_{ji}$	A_i	
(1) High Interference	little prince	ask	0	2.31	2.31	0	3.21	3.21	10.10
(2) Low Interference	you	fall	0.12	3	3.12	0.12	3.9	4.02	0.009
(3) No Interference	he	have	0.21	3	3.21	0.21	3.9	4.11	13.12

(3). For ACT-R-2, the summation of the base-level activation B_i (0.21) and the sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ for the two retrieval cues (3.9) give the overall activation level A_i (4.11), as shown in Table 3.3.

In sum, the lower base-level activation B_i (less recent retrieval) and lower sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ (with matching distractor) for the High Interference configuration give rise to the lower overall activation level A_i. In contrast, the higher A_i for the Low Interference and No Interference configurations comes from both higher B_i (more recent retrieval) and higher $\sum_{j=1}^{m} W_j S_{ji}$ (no matching distractor). The ACT-R model of sentence processing thus predicts higher probability of target retrieval due to higher A_i in the face of more recent retrieval, and in the absence of inhibitory interference, as in the Low and No Interference configurations. When comparing ACT-R-1 with ACT-R-2, the latter receives higher $\sum_{j=1}^{m} W_j S_{ji}$ due to preferential weighting of the structural cues matching the target noun-phrases. Since higher $\sum_{j=1}^{m} W_j S_{ji}$ leads to higher A_i, ACT-R-2 thus predicts higher successful rate of target retrieval compared to ACT-R-1.

Beside the ACT-R metric, we also derive the word-predictability metric by calculating the surprisal value of the critical verbs according to Equation 3.4, and present the values in Table 3.3. The higher the surprisal value, the heavier the cognitive load it requires to process the word. For the three example sentences, the surprisal model thus predicts more processing demand for the High Interference and No Interference configurations, compared to the Low Interference configuration.

3.2.4 Participants

Nineteen Mandarin native speakers (12 female, 6 male, 1 non-binary; mean age=23 years; range 20–38 years) participated in the study. All participants were right-handed, had normal or corrected-to-normal vision and no reported history of neurological disorder. They gave informed consent and were paid 15 USD/hour for their participation. All procedures aligned with protection for human subjects at the University of Michigan, following protocol HUM00081060.

3.2.5 EEG procedure

After debriefing the research procedure, we measured the hearing threshold of every participant per ear by playing 1 kHz tones (300 ms each, 10 ms fade in/out). During the experiment, participants were seated in a chair about 100 cm away from a computer screen in a sound-proof room. They were instructed to listen to the story silently and minimize any movement and eye blinks. The first section of the Chinese "The Little Prince" audiobook was played through the in-ear headphones (EA-2, Etymotic Inc.) at the volume of 45dB above the hearing threshold of the participants. The section lasted for about 10 minutes, and was preceded by a written instruction "the section is about to begin" on the screen. To ensure attentiveness, participants answered four multiple-choice comprehension questions after the section by pressing one of the four arrow keys on the keyboard. All participants exhibited above-average accuracy on the comprehension questions (mean percent correct = 92%, range = [50%, 100%]), with only one participant having 50% accuracy rate due to the smaller number of questions.

3.2.6 EEG recording

Thirty-two actively-amplified electrodes were mounted on an elastic cap (actiCAP, EASYCAP GmbH) and placed on the scalp according to the Standard 32-channel actiCAP snap layout. Bipolar electrodes were placed above and below the left eye to monitor vertical eye movements. The EEG signal was continuously sampled at 500 Hz between 0.1 and 200 Hz, and referenced to the left mastoid electrode. Impedances were maintained at less than 25 kOhms for all electrode sites.

3.2.7 Data analysis

We conducted data processing with the FieldTrip toolbox in MATLAB (Oostenveld et al., 2011). A high-pass filter of 0.1 Hz was first applied, and the data was re-referenced to the average of the left and right mastoid electrodes. Epochs time-locked to the onset of the 223 critical verb from -300–1000 ms were extracted. Afterwards, epochs containing ocular artifact were rejected with Independent Component Analysis (Jung et al., 2000), and epochs containing muscular artifact were removed based on visual inspection. The rejection rate ranged from 0%–4.93% (median = 0.45%) across participants. Epochs from channels above the impedance threshold or with excessive noises were interpolated by surface spline interpolation (Perrin et al., 1987). The number of channels interpolated per participant ranged from 0–3 (median = 0). A 20-Hz low-pass filter was then applied and a 100-ms pre-stimulus baseline was subtracted from all epochs.

For illustration purposes, we divided epochs using a median split on activation level (A_i in Equation 3.1 of the ACT-R model). Specifically, the High Activation group consisted of epochs associated with the 50% highest overall activation level A_i, and the Low Activation group contained epochs with the other 50%. Averaged ERPs were formed from the epochs for plotting purposes only.

To evaluate the predicted sustained negativity effect (Martin et al., 2012; Van Berkum, 2009), we computed single-trial EEG mean amplitude, measured in microvolts, by averaging per trial from central channels (Fz, FC1, FC2, Cz, CP1, CP2, Pz) within three time windows (100–300, 300–500 and 500–800 ms) respectively for each epoch around the critical verb. We then performed Bayesian statistical model analysis with the `brms` package (Bürkner, 2017) and the `loo` package (Vehtari et al., 2017), choosing weakly informative priors to improve convergence and avoid overfitting.[3] The pointwise out-of-sample prediction accuracy from a fitted Bayesian model can be estimated by the Leave-one-out cross-validation (LOO) method with log-likelihood assessed from the posterior simula-

[3] Priors for regression coefficients were defined as $\mathcal{N}(0, 1)$.

tions of the parameter values. This measure of predictive accuracy is called expected log pointwise predictive density (ELPD). Furthermore, the comparison amongst models can be achieved with the estimated difference of expected leave-one-out prediction errors ($\Delta ELPD$) amongst models, as well as the standard error (SE) (Vehtari et al., 2017).

We constructed separate models with the z-scores of single-trial EEG amplitudes as the dependent variable, and the overall activation level A_i, base-level activation B_i, sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$, or GPT2 surprisal as fixed effect, and random slope of each fixed effect by participant. The z-scores of single-trial EEG amplitudes included averages from three separate time windows (100–300, 300–500 and 500–800 ms), and the overall activation level A_i and sum of the weighted associative strength $\sum_{j=1}^{m} W_j S_{ji}$ were derived independently for ACT-R-1 and ACT-R-2.[4]

3.3 Results

For illustration purposes, we divided epochs using a median split on activation level (A_i in Equation 3.1 of the ACT-R model). Specifically, the High Activation group consisted of epochs associated with the 50% highest overall activation level A_i, and the Low Activation group contained epochs with the other 50%. Averaged ERPs were formed from the epochs for plotting purposes only.

Activation level modulated a sustained negativity over anterior electrodes at the critical verb region. This is evident in Figure 3.2 which plots evoked average waveforms for epochs divided by a median-split on activation (computed with ACT-R-1). Figure 3.2 plots the grand average ERPs and 95% confidence intervals at electrodes Fz, Cz, Pz for the High Activation and Low Activation groups in the upper panel, and in the lower panel depicts the corresponding topographic distributions in four adjacent time windows

[4] The models were specified following the Wilkinson-Rogers notation:

1. `eeg` $\sim$ A_i `+ (1 +` A_i `| participant),`
2. `eeg` $\sim$ B_i `+ (1 +` B_i `| participant),`
3. `eeg` $\sim$ $\sum_{j=1}^{m} W_j S_{ji}$ `+ (1 +` $\sum_{j=1}^{m} W_j S_{ji}$ `| participant),`
4. `eeg ~ gpt2_surprisal + (1 + gpt2_surprisal | participant),`

spanning from 0 to 1000 ms after the critical verb onset. The negative shift begins around 100 ms after word onset and lasts to 1000 ms, with a wide scalp distribution. Using EEG, our finding extends previous findings on memory retrieval using functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) (Li et al., 2021).

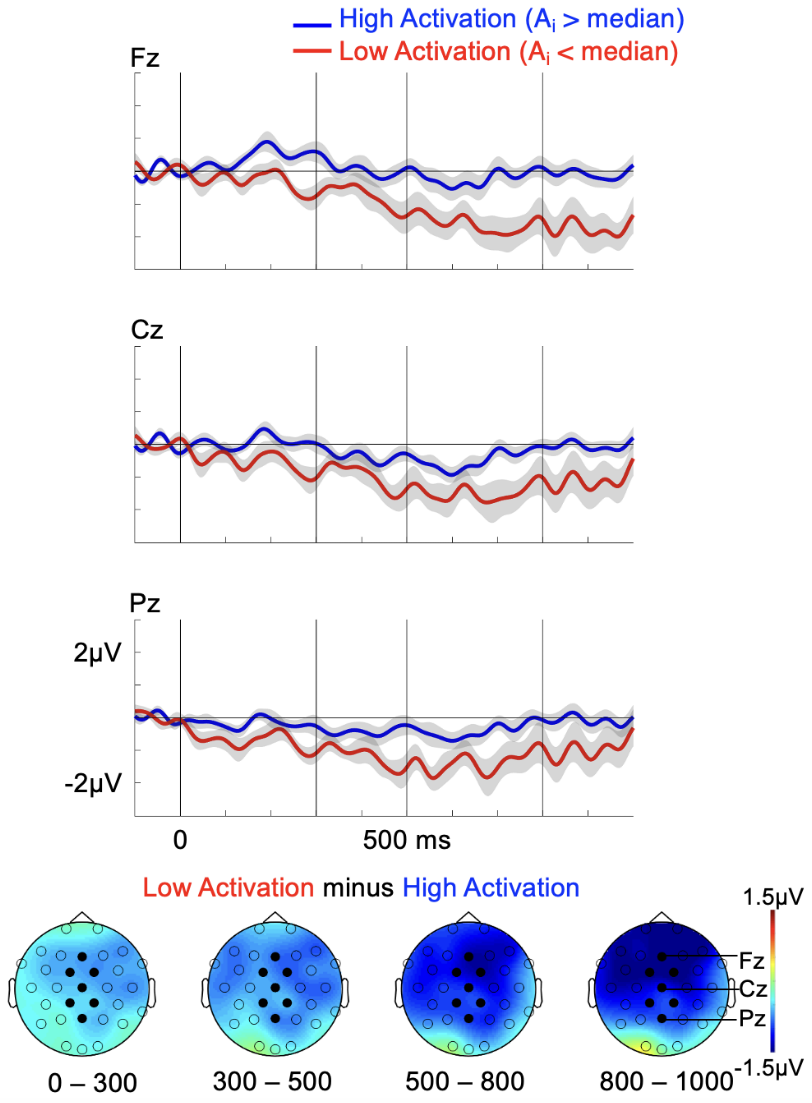

Figure 3.2: Grand averages and 95% confidence intervals (grey shading) at Fz, Cz, Pz for High Activation (blue lines) or Low Activation (red lines) group, and their scalp distributions during four consecutive time windows between 0 to 1000 ms after the critical verb onset.

This pattern was statistically evaluated using Bayesian regression with ACT-R and predictability metrics for each verb and single-trial EEG amplitude. The single-trial Bayesian model analysis illustrated that lower subject activation and weighted associative strength, estimated via both ACT-R models, leads to more negativity in all three time-windows. Table 3.4 presents the estimated mean and 95% Credible Interval (CI) of the posterior distribution of the regression coefficient b for A_i, B_i, $\sum_{j=1}^{m} W_j S_{ji}$ for ACT-R-1 and ACT-R-2 respectively, and surprisal during the 100–300-ms, 300–500-ms and

Table 3.4: The estimated mean and 95% Credible Interval (CI) of the posterior distribution of the regression coefficient b for A_i of ACT-R-1, A_i of ACT-R-2, Surprisal, $\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1, $\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2 and B_i of ACT-R-1&2 during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region.

	100–300 ms			300–500 ms			500–800 ms		
	Estimate	l-95% CI	u-95% CI	Estimate	l-95% CI	u-95% CI	Estimate	l-95% CI	u-95% CI
A_i of ACT-R-1	0.06	0.01	0.10	0.03	-0.01	0.07	0.04	-0.005	0.09
A_i of ACT-R-2	0.06	0.01	0.10	0.03	-0.02	0.07	0.04	-0.005	0.08
Surprisal	0.01	-0.02	0.04	-0.002	-0.04	0.03	-0.002	-0.03	0.03
$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1	0.04	0.004	0.07	0.04	0.007	0.08	0.04	0.006	0.07
$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2	0.04	0.002	0.07	0.04	0.007	0.08	0.04	0.008	0.07
B_i of ACT-R-1&2	0.03	-0.001	0.07	0.003	-0.03	0.04	0.02	-0.02	0.05

500–800-ms time windows at the critical verb region.

In the next step, we directly compared all models specified in Footnote 4 using approximate leave-one-out cross-validation (Vehtari et al., 2017). The comparison can be summarized by the difference in expected log pointwise predictive density ($\Delta ELPD$, see Section 3.2.7 for using $\Delta ELPD$ as a tool for model comparison). To begin with, we found that compared to GPT2 surprisal, ACT-R metrics received stronger evidence for successfully predicting single-trial EEG amplitude of the sustained negativity, and that the ACT-R-1 and ACT-R-2 models showed comparable performance. Table 3.5 depicts the $\Delta ELPD$ and standard error (SE) for the model comparison among A_i of ACT-R-1, A_i of ACT-R-2, and Surprisal during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region.

Secondly, the better predictive power of A_i could be attributed to its sub-component, $\sum_{j=1}^{m} W_j S_{ji}$, but not B_i, in both ACT-R-1 and ACT-R-2 models. For ACT-R-1, Table 3.6 displays the $\Delta ELPD$ and SE for the model comparison among A_i, B_i, $\sum_{j=1}^{m} W_j S_{ji}$, and Surprisal during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region And Table 3.7 displays similar comparison using the metric values of ACT-R-2.

We thus report one of the first cortical electrophysiological evidence of the memory interference effects during naturalistic language processing, and suggest that interference modeled with ACT-R may generalize to more everyday comprehension situation.

Table 3.5: The difference in expected log pointwise predictive density ($\Delta ELPD$) and standard error (SE) for the model comparison among A_i of ACT-R-1, A_i of ACT-R-2, and Surprisal during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region.

	100–300 ms			300–500 ms			500–800 ms	
	$\Delta ELPD$	SE		$\Delta ELPD$	SE		$\Delta ELPD$	SE
A_i of ACT-R-2	0.0	0.0	A_i of ACT-R-1	0.0	0.0	A_i of ACT-R-1	0.0	0.0
A_i of ACT-R-1	0.0	0.1	A_i of ACT-R-2	0.0	0.2	A_i of ACT-R-2	0.0	0.2
Surprisal	-2.7	2.6	Surprisal	-0.9	1.6	Surprisal	-1.5	2.3

Table 3.6: The difference in expected log pointwise predictive density ($\Delta ELPD$) and standard error (SE) for the model comparison among A_i of ACT-R-1, B_i of ACT-R-1, $\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1, and Surprisal during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region.

	100–300 ms			300–500 ms			500–800 ms	
	$\Delta ELPD$	SE		$\Delta ELPD$	SE		$\Delta ELPD$	SE
A_i of ACT-R-1	0.0	0.0	$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1	0.0	0.0	$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1	0.0	0.0
$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-1	-0.4	1.9	A_i of ACT-R-1	-2.9	2.7	A_i of ACT-R-1	-1.2	1.8
B_i of ACT-R-1	-0.8	1.2	B_i of ACT-R-1	-3.5	3.7	B_i of ACT-R-1	-2.6	2.7
Surprisal	-2.7	2.6	Surprisal	-3.7	3.6	Surprisal	-2.7	3.0

Table 3.7: The difference in expected log pointwise predictive density ($\Delta ELPD$) and standard error (SE) for the model comparison among A_i of ACT-R-2, B_i of ACT-R-2, $\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2, and Surprisal during the 100–300-ms, 300–500-ms and 500–800-ms time windows at the critical verb region.

	100–300 ms			300–500 ms			500–800 ms	
	$\Delta ELPD$	SE		$\Delta ELPD$	SE		$\Delta ELPD$	SE
A_i of ACT-R-2	0.0	0.0	$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2	0.0	0.0	$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2	0.0	0.0
$\sum_{j=1}^{m} W_j S_{ji}$ of ACT-R-2	-0.2	1.9	A_i of ACT-R-2	-2.9	2.8	A_i of ACT-R-2	-1.1	1.8
B_i of ACT-R-2	-0.8	1.2	B_i of ACT-R-2	-3.5	3.8	B_i of ACT-R-2	-2.5	2.7
Surprisal	-2.7	2.6	Surprisal	-3.8	3.7	Surprisal	-2.6	3.0

3.4 Discussion

We test for the effects of memory retrieval (i.e., interference and decay) and predictability and their neural bases during the resolution of subject-verb agreement during Chinese audiobook naturalistic listening. Extending Li et al. (2021)'s characterisation of pronoun reference in natural stories with ACT-R, we fit electroencephalography (EEG) signals against two ACT-R variants (Lewis and Vasishth, 2005) and word-predictability measure from Chinese GPT2 (Du, 2019; Radford et al., 2019). Variants differ in cue-combinatorics (Parker and Phillips, 2017; Sturt, 2003): structural cues may be weighted (i) equally (Lewis and Vasishth, 2005; McElree, 2000; Vasishth et al., 2008) or (ii) preferably (Parker and Phillips, 2017; Van Dyke and McElree, 2011; Yadav et al., 2022) over non-structural cues in ACT-R.

We first successfully report the effects of memory retrieval during naturalistic comprehension. Our EEG results showed that a sustained negativity was elicited by Low Activation vs. High Activation group. This negativity was better captured by both ACT-R models, but not the word-predictability model. Within the ACT-R metric, we further identify the interference component as the driving force behind this negativity effect, as compared to the decay component in the ACT-R model.

Theoretically, our results support the general cue-based working-memory retrieval theory (Lewis and Vasishth, 2005; Vasishth et al., 2008), but not the noisy-memory-based-predictability account (Ryu and Lewis, 2021). In order to further tease apart sub-theories concerning cue-preference under the cue-based retrieval framework, and to probe the potential interaction between memory retrieval and predictability, more future work is needed. Firstly, to test for the cue-preference-by-predictability account (Parker and Phillips, 2017; Tung and Brennan, 2023), cue-weighting could be varied in a more continuous fashion. Rather than assigning an invariant weighting (e.g., 1 or 1.6) for the retrieval cues for all target words, one strategy would be to parameterize the weighting scheme per word according to the predictability of the critical words at the retrieval site. This will provide a novel dynamic metric for evaluating the influence of cue-weighting on the memory retrieval process for each individual target word. This approach also constitutes efforts to more tightly integrate memory retrieval and predictability into a common processing model. Secondly, to assess the cue-preference-by-proficiency account (Yadav et al., 2022), the language proficiency of participants could be estimated by separate battery of tests, which will in turn affect the cue-weighting scheme for each participant. Taking into account individual differences could further improve the predictive power of the memory metric. Finally, to evaluate the cue-preference-by-morphosyntax account (Mertzen et al., 2023), cross-linguistic comparison within subject might better represent the cue-weighting scheme at work in the individual minds of each participant. For example, comparing the metric performance on explaining data on English audiobook listening to data on German audiobook listening might be a valuable benchmark for theories/metrics built previously on controlled experiments.

Methodologically, we also illustrated how to combine computational modeling and cognitive neuroscience to study continuous spoken language comprehension in real time. Specifically, we exploited both symbolic modeling and large language models in combination with event-related brain potential recordings from the human scalp to show that the incremental parser exploited both syntactic and semantic information to retrieve target noun-phrases from memory.

3.5 Conclusion

Our study connects theories of memory retrieval and prediction with their neural implementations, suggesting a strong correlation between cue-based retrieval mechanism and the amplitude of a sustained negativity ERP component in response to processing the verb during subject-verb dependency while listening to an audiobook story. These computationally explicit theories generate values as linking hypotheses to test against measured human responses, and are therefore especially useful in decoding EEG data from naturalistic studies with non-factorial design and naturally occurring sentences as stimuli. We thus demonstrate how the formal modeling approach could simulate cognitive processes in language comprehension that is generalizable to every-day language use.

CHAPTER IV

Conclusion

This dissertation investigates how predictions and working memory demands affect word-by-word language processing, and identifies the neural mechanisms that mediate these effects. These experiments, based on Mandarin Chinese, complement existing work on typologically different languages (primarily English and German), and thus serves to broaden the cross-linguistic data contributing to theories of memory and prediction under analysis. To integrate linguistic predictions (Hale, 2001; Levy, 2008) and working memory load (Lewis et al., 2006; McElree, 2006; Vasishth et al., 2019) into a unified model of processing complexity, Chapter II tests for and reports, for the first time, empirical evidence for the modulation of memory retrieval by linguistic predictions using controlled experiments. Chapter III then isolates, also for the first time, the memory effects independently from the predictability effects using an authentic audiobook story during continuous speech comprehension. Below I will summarize the contributions of Chapters II and III before comparing diverse neural signatures of memory retrieval. Future directions pertaining to the refinement of memory and predictability models, alongside cross-linguistic comparison will also be discussed.

Chapter II deploys a within-task design (i.e. sentence-reading) to assess whether Mandarin constructions show an interaction between predictability and memory interference, as previously demonstrated for English. This EEG experiment focuses on the "noun-phrase ellipsis" (e.g., Elizabeth ate a **cookie** that was next to the **cake** and Harriet also ate **<u>one</u>**.). Similar to prior work in English, I found a greater positive voltage over the centro-posterior scalp areas around 600 milliseconds after the critical word onset. This "P600 effect" has been proposed to index processing difficulty due to retrieval interference in ungrammatical sentences (Tanner et al., 2014; Xiang et al., 2009). Crucially, this facilitatory interference effect only surfaced when the target item was highly predictable, but not when the target and distractor items were equally unpredictable. The results support my hypothesis and align with a working memory cost account that incorporates predictability (Parker and Phillips, 2017). In other words, these neuro-electrical dynamics are consistent with an integrated perspective where linguistic predictions and memory load jointly and interactively contribute to successful language comprehension (Futrell et al., 2020). The memory-based (Lewis et al., 2006; McElree, 2006; Vasishth et al., 2019) and prediction-based (Hale, 2001; Levy, 2008) accounts are, therefore, not mutually exclusive, but may rather represent different aspects of the processing mechanisms.

Chapter III capitalizes on a naturally occurring Mandarin audiobook story to investigate whether the memory retrieval and predictability effects also surface in a more naturalistic environment of language comprehension. This EEG experiment studies subject-verb dependencies (e.g., **You** for **me** <u>**draw**</u> a sheep.). In line with previous English literature, I discovered an increased negative voltage with wide scalp distribution starting around 200 ms after the critical word onset. This "Nref" effect (Van Berkum, 2009; Van Berkum et al., 1999) has been associated with processing difficulty due to retrieval interference in grammatical contexts (Martin et al., 2012). Furthermore, this inhibitory interference effect (Jäger et al., 2017, for review) was better explained by the ACT-R memory (Lewis and Vasishth, 2005) model as compared to the GPT-2 probability model (Ryu and Lewis, 2021). I thus reported one of the first cortical electrophysiological evidence for how memory retrieval is modulated during naturalistic language processing.

4.1 Neural signatures of memory retrieval

One nuance of the experimental results is that memory retrieval effects are indexed by the P600 ERP component at the critical word region in Chapter II, but by the Nref component in Chapter III. The critical word region is the retrieval site for the target word in the process of dependency resolution. It corresponds to the word region of number-classifier sequence for noun-phrase ellipsis construction, and to the region of verb for subject-verb dependency. Importantly, in Chapter II, while the interference effect correlates with the P600 component at the critical word region, the same interaction effect is associated with the Nref component at the word region immediately following the critical word region. The presence of the P600 effect and the distinct timing of the Nref effects might relate to several differences in the two experiments. Chapter II employs a reading task with carefully designed individual ungrammatical sentences. But in Chapter III, participants listen to a naturally occurring audiobook story that contains only grammatical sentences.

First, the two experiments differ in modality. Second, and perhaps more importantly, the experiment in Chapter II focuses on memory effects in ungrammatical sentences (facilitatory interference) while the naturalistic analysis in Chapter III focuses on ERPs for grammatical sentences. Since the P600 effect has been consistently elicited by violations of syntactic principles (Hagoort et al., 1993; Kaan et al., 2000; Kaan, 2002; Kaan and Swaab, 2003; Molinaro et al., 2011; Tanner et al., 2014; Yang et al., 2015), as well as by facilitatory interference under ungrammatical contexts (Tanner et al., 2014; Xiang et al., 2009), it is predicted for the ungrammatical stimuli sentences in Chapter II, but not necessarily for the grammatical audiobook texts in Chapter III.

The Nref effect, on the other hand, has been induced when a unique referent cannot be selected from several competitors given prior context (Van Berkum et al., 1999; Van Berkum, 2009), as well as when distractors cause interference in grammatical situations (Martin et al., 2012). A statistically non-significant trend of a facilitatory interference effect in ungrammatical sentences indexed by the Nref effect was also reported by Martin et al. (2012). Not strongly tied with the grammaticality manipulation, the Nref effect can thus be predicted for stimuli introducing referential processing difficulty in both grammatical and ungrammatical situations, as in both of our experiments. Specifically, due to distractor interference, a sufficiently unique target noun-phrase antecedent might become difficult to be identified in the noun-phrase ellipsis construction in Chapter II. Similar difficulty holds for the target subject noun-phrase in the subject-verb dependency in Chapter III in the presence of distractor interference. The delayed timing of the Nref effect in Chapter II compared to Chapter III might reflect the additional immediate effort required to resolve syntactic violation before the referential ambiguity.

4.2 Modeling naturalistic comprehension

The findings of Chapter III further complement the results of Li et al. (2021), who also successfully characterized brain activities during story listening using the ACT-R memory model, but not deep neural network models with LSTM or Transformer architecture. They identified a cortical network involving the anterior and posterior left middle temporal gyrus and the angular gyrus for pronoun resolution using fMRI and MEG experiments. And they attributed the underlying cognitive processes for referential processing to domain-general mechanism similar to memory retrieval. I extended this line of research to a more diverse set of linguistic constructions and experimental methodologies. Namely, I used subject-verb dependency in an EEG experiment to provide converging evidence in support of the underlying mechanisms for online language processing that is generalizable to every-day situation. To further understand the detailed memory retrieval operations, I additionally evaluated several theories of cue-preference by comparing two ACT-R models.

By pitching the memory model directly against the predictability model, I successfully isolated the memory retrieval effects from linguistic predictability effects by identifying unique neural correlates of the memory demand. While the current study could not further tease apart sub-theories of cue-preference under the general cue-based memory retrieval framework, I provide feasible ways forward to try to test for those delicate theoretical nuances. Specifically, I suggest to: (i) parameterize cue-weighting with a continuous scale according to the predictability of the critical words in order to test for the cue-preference-by-predictability account (Parker and Phillips, 2017; Tung and Brennan, 2023), (ii) measure language proficiency across native speakers and vary cue-weighting accordingly to evaluate the cue-preference-by-proficiency account (Yadav et al., 2022), and (iii) collect data on English and German audiobook listening to assess the cue-preference-by-morphosyntax account (Mertzen et al., 2023). These approaches serve to investigate the hypothesized individual differences among words, participants and language stimuli, with the goal of maximizing the capabilities of the memory metrics in approximating human performance on naturalistic language comprehension.

4.3 Cross-linguistic comparison

While my dissertation investigates the effect of memory and predictability using both artificially constructed Mandarin single-sentences stimuli and more ecologically valid Mandarin audiobook stimuli on native (L1) speakers, the next natural step is to use English audiobook stimuli on second-language (L2) learners to further achieve cross-competency and cross-linguistic comparison. While theories of sentence processing and cognitive architecture flourish with growing clarity, they are severely constrained by a small class of language users. The next urgent step in the field is to connect the models with a more representative group of languages and learners. One possible extension of the research on naturalistic comprehension is to compare L1 and L2 speakers of English and that of Mandarin. This dynamic extension seeks to unveil the cognitive and neural mechanisms underlying multilingual processing success/difficulty, especially since the engagement of predictive mechanism might vary as a function of linguistic typology or proficiency (Blasi et al., 2022; Huettig and Mani, 2016). Apart from further data collection, continued refinement of formal computational models of memory and predictability (Hahn et al., 2022; Lewis and Vasishth, 2005; Parker and Phillips, 2017) will benefit the aims of building biologically plausible and interpretable models for human electrophysiological data.

Pushing forward, I am keen to extend this line of research to indigenous languages, which are underrepresented in neurolinguistics studies. Experimental data is scarce (cf. Wagers et al. (2018)), and data collection poses unique challenges. My naturalistic audiobook listening paradigm tackles the challenges by offering an accessible approach for populations (e.g., non-literate) for whom traditional experimental tasks may be difficult. Investigating the parsing strategies for diverse languages could unveil the universality and variance among linguistic algorithms, their neural signature, and downstream effects of comprehension success/breakdown, which hold both educational and therapeutic implications.

APPENDIX

APPENDIX A

Supplementary Material of Chapter II

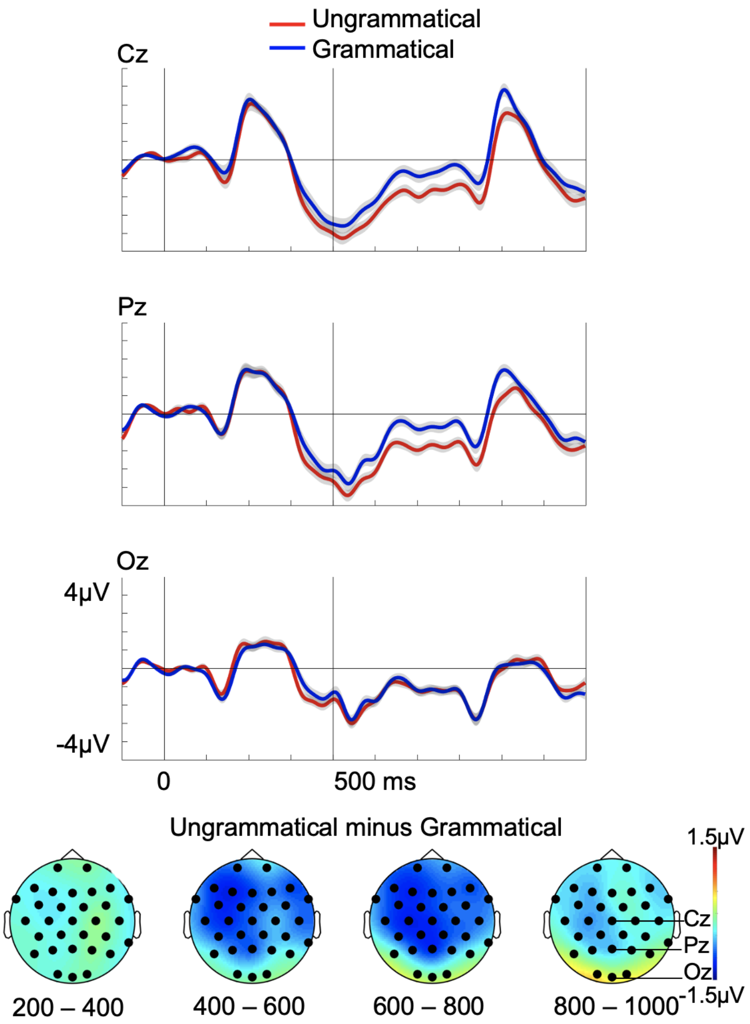

Figure A.1: Grand averages and 95% confidence intervals (grey shading) at Cz, Pz, Oz for Ungrammatical (red lines) or Grammatical (blue lines) condition, and their scalp distributions during four consecutive time windows between 200 to 1000 ms after the verb onset immediately following the critical number-classifier sequence.

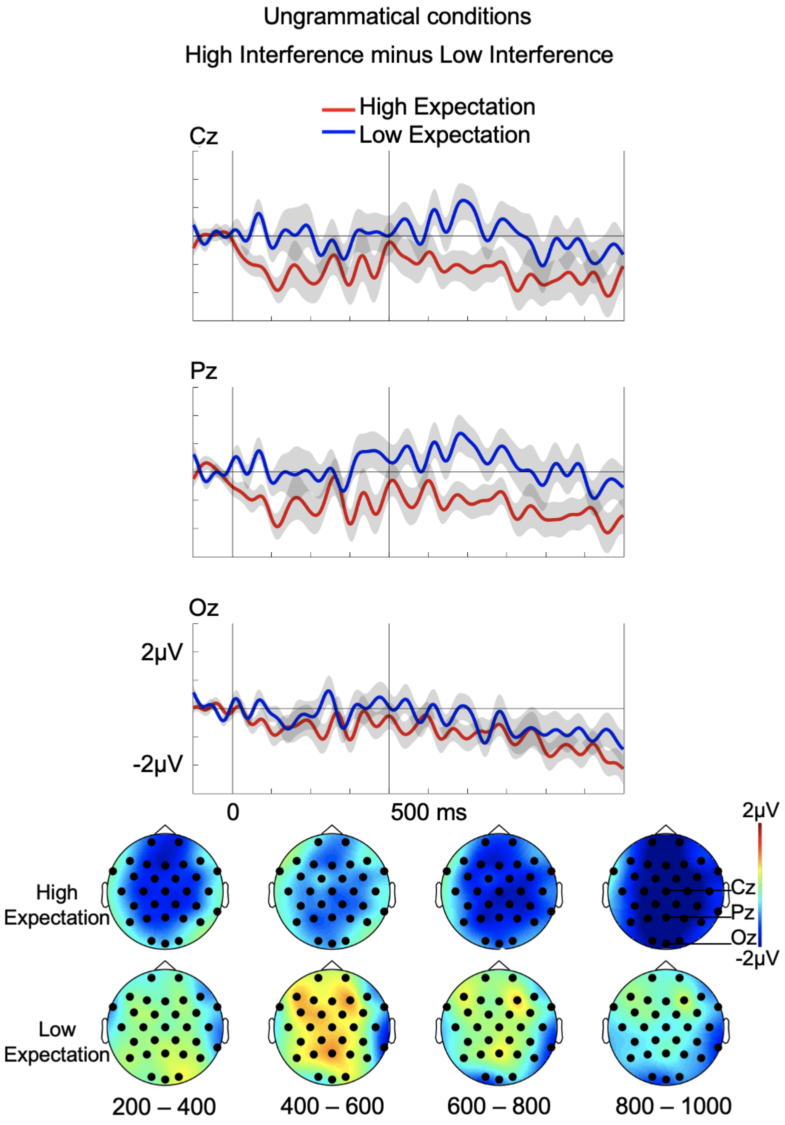

Figure A.2: Difference waves and 95% confidence intervals (grey shading) at Cz, Pz, Oz for High Interference minus Low Interference in High Expectation (red lines) or Low Expectation (blue lines) ungrammatical condition, and their scalp distributions during four consecutive time windows between 200 to 1000 ms after the verb onset immediately following the critical number-classifier sequence.

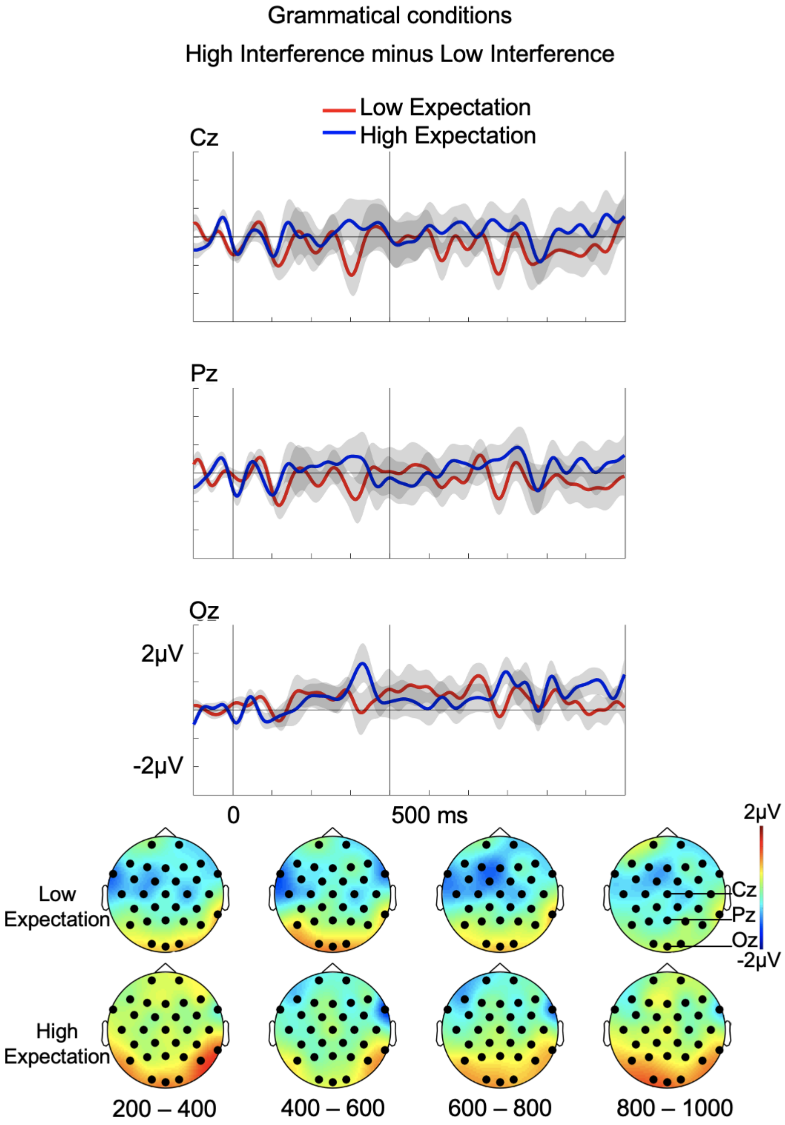

Figure A.3: Difference waves and 95% confidence intervals (grey shading) at Cz, Pz, Oz for High Interference minus Low Interference in Low Expectation (red lines) or High Expectation (blue lines) grammatical condition, and their scalp distributions during four consecutive time windows between 200 to 1000 ms after the verb onset immediately following the critical number-classifier sequence.

BIBLIOGRAPHY

BIBLIOGRAPHY

Allan, K. (1977). Classifiers. *Language*, 53(2):285–311.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Psychology Press.

Arnett, N. and Wagers, M. (2017). Subject encodings and retrieval interference. *Journal of Memory and Language*, 93:22–54.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., and Majid, A. (2022). Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170.

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., and Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and language*, 117(3):133–152.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1).

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Brennan, J. R. and Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., and Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Brothers, T. and Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Campanelli, L., Van Dyke, J. A., and Marton, K. (2018). The modulatory effect of expectations on memory retrieval during sentence comprehension. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, page 1434“1439, Austin, Texas. Cognitive Science Society.

Chan, S.-h. (2019). An elephant needs a head but a horse does not: An erp study of classifier-noun agreement in mandarin. *Journal of Neurolinguistics*, 52:100852.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Chen, Z. and Hale, J. T. (2021). Quantifying structural and non-structural expectations in relative clause processing. *Cognitive Science*, 45(1):e12927.

Cheng, L. L.-S. and Sybesma, R. (2014). Syntactic structure of noun phrases. In Huang, J. C.-T., Li, A. Y.-H., and Simpson, A., editors, *The Handbook of Chinese Linguistics*, pages 248–274. John Wiley & Sons.

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

Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45(2):145–171.

Cunnings, I. and Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language*, 102:16–27.

Del Gobbo, F. (2014). Classifiers. In Huang, J. C.-T., Li, A. Y.-H., and Simpson, A., editors, *The Handbook of Chinese Linguistics*, pages 26–48. John Wiley & Sons.

DeLong, K. A., Quante, L., and Kutas, M. (2014). Predictability, plausibility, and two late erp positivities during written sentence comprehension. *Neuropsychologia*, 61:150–162.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Du, Z. (2019). Gpt2-chinese: Tools for training gpt2 model in chinese language. https://github.com/Morizeyao/GPT2-Chinese.

Engelmann, F., Jäger, L. A., and Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, 43(12):e12800.

Federmeier, K. D. and Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research*, 8(3):373–392.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., and Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain research*, 1146:75–84.

Franck, J., Colonna, S., and Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in psychology*, page 349.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Friederici, A. D., Mecklinger, A., Spencer, K. M., Steinhauer, K., and Donchin, E. (2001). Syntactic parsing preferences and their on-line revisions: A spatio-temporal analysis of event-related brain potentials. *Cognitive Brain Research*, 11(2):305–323.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Gao, M. Y. and Malt, B. C. (2009). Mental representation and cognitive consequences of chinese individual classifiers. *Language and Cognitive Processes*, 24(7-8):1124–1179.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., and Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain and language*, 126(3):314–326.

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the p600. *Language and cognitive processes*, 25(2):149–188.

Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and cognitive processes*, 8(4):439–483.

Hahn, M., Futrell, R., Levy, R., and Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. R. (2018). Finding syntax in human encephalography with beam search. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1:2727“2736.

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

Henderson, J. M., Choi, W., Lowder, M. W., and Ferreira, F. (2016). Language structure in the brain: A fixation-related fmri study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

Hsiao, F. and Gibson, E. (2003). Processing relative clauses in chinese. *Cognition*, 90(1):3–27.

Hsu, C.-C., Tsai, S.-H., Yang, C.-L., and Chen, J.-Y. (2014). Processing classifier–noun agreement in a long distance: An erp study on mandarin chinese. *Brain and Language*, 137:14–28.

Huang, C.-R. (2009). Tagged chinese gigaword version 2.0, ldc2009t14. *Linguistic Data Consortium*.

Huang, C.-R., Kilgarriff, A., Wu, Y., Chiu, C.-M., Smith, S., Rychlỳ, P., Bai, M.-H., and Chen, K.-J. (2005). Chinese sketch engine and the extraction of grammatical collocations. In *Proceedings of the fourth SIGHAN workshop on Chinese language processing*.

Huettig, F. and Mani, N. (2016). Is prediction necessary to understand language? probably not. *Language, Cognition and Neuroscience*, 31(1):19–31.

Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., and Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111:104063.

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., and Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, 111(10):1745–1758.

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An erp study. *Journal of Psycholinguistic Research*, 31(2):165–193.

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The p600 as an index of syntactic integration difficulty. *Language and cognitive processes*, 15(2):159–201.

Kaan, E. and Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of cognitive neuroscience*, 15(1):98–110.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, 2(4):647–670.

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

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602.

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

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

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of memory and language*, 82:18–40.

Kush, D., Lidz, J., and Phillips, C. (2017). Looking forwards and backwards: The real-time processing of strong and weak crossover. *Glossa (London)*, 2(1).

Kush, D. and Phillips, C. (2014). Local anaphor licensing in an sov language: Implications for retrieval strategies. *Frontiers in psychology*, 5:1252.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp). *Annual review of psychology*, 62:621“647.

Kutas, M., Federmeier, K. D., and Urbach, T. P. (2014). The “negatives” and “positives” of prediction in language. In Michael S. Gazzaniga, G. R. M., editor, *The Cognitive Neurosciences*, page 649“656. MIT Press.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2015). Agreement attraction in spanish comprehension. *Journal of Memory and Language*, 82:133–149.

Lau, E., Stroud, C., Plesch, S., and Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and language*, 98(1):74–88.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In van Gompel R. P. G., editor, *Sentence Processing*, pages 78–114. Psychology Press.

Levy, R. P. and Keller, F. (2013). Expectation and locality effects in german verb-final structures. *Journal of memory and language*, 68(2):199–222.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29:375–419.

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10):447–454.

Li, A. Y.-H. and Wei, T.-C. (2014). Ellipsis. In Huang, J. C.-T., Li, A. Y.-H., and Simpson, A., editors, *The Handbook of Chinese Linguistics*, pages 275–310. John Wiley & Sons.

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., and Hale, J. (2022). Le petit prince multilingual naturalistic fmri corpus. *Scientific data*, 9(1):1–15.

Li, J., Wang, S., Luh, W.-M., Pylkkänen, L., Yang, Y., and Hale, J. (2021). Cortical processing of reference in language revealed by computational models. *BioRxiv*, pages 2020–11.

Logačev, P. and Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, 40(2):266–298.

Lopopolo, A., Frank, S. L., Van den Bosch, A., and Willems, R. M. (2017). Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PloS one*, 12(5):e0177794.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

Martin, A. E. (2018). Cue integration during sentence comprehension: Electrophysiological evidence from ellipsis. *PloS one*, 13(11):e0206616.

Martin, A. E. and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(5):1231.

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage*, 59(2):1859–1869.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, 29(2):111–123.

McElree, B. (2006). Accessing recent events. In Ross, B. H., editor, *Psychology of Learning and Motivation*, volume 46, pages 155–200. Elsevier, San Diego, CA.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91.

Merchant, J. et al. (2001). *The syntax of silence: Sluicing, islands, and the theory of ellipsis*, volume 1. Oxford University Press on Demand.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., and Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from english and german eye-tracking data. *Glossa Psycholinguistics*, 2(1).

Molinaro, N., Barber, H. A., and Carreiras, M. (2011). Grammatical agreement processing in reading: Erp findings and future directions. *cortex*, 47(8):908–930.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, 10(5-6):389–395.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., et al. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, 375(1791):20180522.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, 7:e33468.

Oh, B.-D., Clark, C., and Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Oh, B.-D. and 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.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011.

Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806.

Parker, D. and Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, 94:272–290.

Perrin, F., Pernier, J., Bertnard, O., Giard, M.-H., and Echallier, J. (1987). Mapping of scalp potentials by surface spline interpolation. *Electroencephalography and clinical neurophysiology*, 66(1):75–81.

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

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 324–333.

Ryu, S. H. and Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *arXiv preprint arXiv:2104.12874*.

Schoknecht, P., Roehm, D., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2022). The interaction of predictive processing and similarity-based retrieval interference: an erp study. *Language, Cognition and Neuroscience*, pages 1–19.

Sivula, T., Magnusson, M., and Vehtari, A. (2020). Uncertainty in bayesian leave-one-out cross-validation based model comparison. *arXiv preprint arXiv:2008.10296*.

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., and Hale, J. T. (2023). Modeling structure-building in the brain with ccg parsing and large language models. *Cognitive Science*, 47(7):e13312.

Stolz, W. S. (1967). A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior*, 6(6):867–873.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3):542–562.

Tanner, D., Nicol, J., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of memory and language*, 76:195–215.

Tucker, M. A., Idrissi, A., and Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from arabic. *Frontiers in psychology*, 6:347.

Tung, T.-Y. and Brennan, J. R. (2023). Expectations modulate retrieval interference during ellipsis resolution. *Neuropsychologia*, page 108680.

Van Berkum, J. J. (2009). The neuropragmatics of'simple'utterance comprehension: An erp review. In Sauerland, U. and Yatsushiro, K., editors, *Semantics and pragmatics: From experiment to theory*, pages 276–316. Palgrave Macmillan, Basingstoke.

Van Berkum, J. J., Brown, C. M., and Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of memory and language*, 41(2):147–182.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407.

Van Dyke, J. A. and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Van Dyke, J. A. and McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of memory and language*, 55(2):157–166.

Van Dyke, J. A. and McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of memory and language*, 65(3):247–263.

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712.

Vasishth, S. and Drenhaus, H. (2011). Locality in german. *Dialogue & Discourse*, 2(1):59–82.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11):968–982.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Wagers, M. W., Borja, M. F., and Chung, S. (2018). Grammatical licensing and relative clause parsing in a flexible word-order language. *Cognition*, 178:207–221.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Wicha, N. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of cognitive neuroscience*, 16(7):1272–1288.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: Erp evidence. *Brain and Language*, 108(1):40–55.

Xiang, M. and Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, cognition and neuroscience*, 30(6):648–672.

Xu, L., Zhang, X., and Dong, Q. (2020). Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv*, abs/2003.01355.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., and Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate bayesian computation. *Open Mind*, 6:1–24.

Yang, C.-L., Perfetti, C. A., and Liu, Y. (2010). Sentence integration processes: An erp study of chinese sentence comprehension with relative clauses. *Brain and Language*, 112(2):85–100.

Yang, Y., Wu, F., and Zhou, X. (2015). Semantic processing persists despite anomalous syntactic category: Erp evidence from chinese passive sentences. *PloS one*, 10(6):e0131936.

Zhang, N. N. (2009). *Coordination in syntax*, volume 123. Cambridge University Press.

Zhang, Y., Zhang, J., and Min, B. (2012). Neural dynamics of animacy processing in language comprehension: Erp evidence from the interpretation of classifier–noun combinations. *Brain and Language*, 120(3):321–331.