Incremental structure prediction during language comprehension:

Behavioral and neurobiological evidence from Mandarin Chinese and English

Integrated Thesis for the Degree of D.Phil in Linguistics, Philology and Phonetics

University of Oxford

Michaelmas Term, 2025

Oriel College

Candidate: Zirui Huang

Word count: 51057

ACKNOWLEDGEMENTS

I imagined this moment countless times while I was wrestling with my thesis. I often wondered what it would feel like to look back on my five years in Oxford. Now, my first thought is, it has been an unforgettable journey—but nothing like the one I once pictured. People always say that doing a DPhil is a lonely journey, but I believed I would be fine: the thrill of imagining a colourful and fulfilling future far outweighed any fear of solitude. Then Covid-19 swept in alongside the start of my DPhil, throwing everyone into a period of ultimate loneliness. My plans fell apart. My expectations collapsed. I suffered from losses and regrets. Many times, I felt overwhelmed by the vast uncertainty ahead of me, like a plane hurtling into a thick bank of clouds, losing all sense of direction. But luckily, I was not without guidance. I could still see lights on the ground, guiding me toward a safe landing and giving me courage for the next adventure. This acknowledgement is my chance to formally express my gratitude to those "lights" in my little airport.

First and foremost, I am grateful for that one thing that never fails to amaze me and keeps my passion for research alive, the complexity of human language. Year after year exploring how our brain makes sense of language, I do not feel a sense of achievement but feel humble and eager to learn more. Language works on its fuzziness and thus studying it can be quite tricky, but not a single day have I felt that it was pointless to keep digging into how language works in our brain. There is nothing more I could ask for.

It is also sometimes tempting to borrow a language-processing perspective when thinking about life in the real world. One fundamental observation in psycholinguistics is that sentence processing is remarkably incremental: no matter how hard it is to keep parsing, be it ambiguity or lack of information, our brain always tries its best to achieve full

incrementality, meaning making immediate syntactic decisions for every incoming word. Of course, in this way, our brain could make bad choices and the sentence needs to be reanalyzed. Normally, it is the moment when fear crepts into our minds: can we avoid making mistakes and being forced to start things all over again? What if we wait and see further? What if we make our decision later? But no, our brain never waits. Our brain always chooses to "go for it". Even if it makes a wrong decision, it will just recover and reconstruct. It seems like we could grasp more certainty if we delay our choices, but our brain has already decided for us that it is more important to keep our sharp intuition and determination to move things forward. So every time I am hesitant and overwhelmed by the uncertainties and choices in life, I always remind myself: just pick one direction and go. Even if we are wrong or at some point we regret, we can always recover and move forward again.

Then of course, my deeply heartfelt thanks go to my supervisor Matt Husband, who step by step "babysits" me into the world of psycholinguistics. I can still recall our first talk about my MPhil topic, and I saw passion in his eyes when talking about what we could potentially do. From that moment on, psycholinguistics was not something that looked cool, it became something REALLY cool. Every time I overthink too much, he is always here showing me the way out of my messy thoughts. Every time I become worried about the uncertainties of the experiments, he taught me that "it is the fun part of doing research!" I will always remember the excitement of working with Matt and carry on doing my research on "something sexy".

I am also grateful for all the support from our Linguistic Faculty, especially Prof. Wolfgang de Melo, Prof. Aditi Lahiri, and Prof. Hanne Eckhoff, for their patience in

advising me on my works in typology, psycholinguistics, and syntax. Their guidance shaped my views of linguistic research.

I am always eager, maybe a bit over-eager, to explore interdisciplinary possibilities in linguistic research and to learn new techniques. I feel extremely lucky and grateful that I have met so many professors willing to host me for academic visits and work opportunities, and teach me new skills. I owe a great debt of gratitude to Prof. Qingqing Qu for her supervision of my first and second EEG experiments. Her trust and support for me to conduct my research during the pandemic meant everything to me at that time. I am deeply thankful to Prof. Shravan Vasishth for enhancing my statistics skills for data processing and encouraging me to explore the world of computational modeling. I am sincerely grateful to my current boss, Prof. Alexis Hervais-Adelman, for giving me the opportunity to work on the ProPoSAL project using MEG data and for opening up to me the perspective of neuroscience research. His passion for cutting-edge methods in computational neuroscience has also inspired me to challenge myself and expand my skillset to machine learning.

I would like to express my deepest gratitude to my thesis examination committee, Prof. Gillian Ramchand and Prof. Masaya Yoshida, for their insightful feedback and generous support throughout the process of evaluating my thesis. The viva was an immensely rewarding experience that helped me think more deeply about how to present my research clearly and convincingly, and it has inspired many new ideas for my future work. I did not feel that I was being examined; instead, it felt like an engaging and thought-provoking conversation about my DPhil research journey.

I want to give my heartfelt thanks to all the participants for my experiment. It means so much to me when many of them told me that they thought my experiment was interesting and they were very curious about my study. I worried a lot about what if people thought my experiment was boring and no one took it seriously, or even worse, what if no one came to my experiment. But each one of my participants gradually built up my confidence when they said they liked it and kindly offered to introduce their friend to do my experiment. Probably they never knew that they were the motivation to push me through the tiring and worrying period of data collection, but I have to say, thank you for your participation, sincerely!

Now it's time for some more personal notes for my family and friends. I cannot imagine my life without them. Of course, I start with my parents. I am really grateful that they gave me the freedom to build up my own life and taught me to take responsibility for my own decisions. It is a blessing that my mom always tries to understand me and respect me as an adult. And my dad has been patiently enduring two emotional and talkative women under one roof. Although I never say it out, I can never go this far without their unconditional love for me.

Then I'd like to thank Jan Dobler, for being such a supportive and caring partner. For a helpless procrastinator like me, he has dutifully acted as my taskmaster, keeping me on track. When I feel annoyed, he can always find ways to calm me down. When I feel discouraged, he can find even more ways to make me laugh. Thanks to him, the time when I had to juggle my work and my thesis became much more bearable.

My life in Oxford would not be so colorful without my friends Songjun He, Rebecca Zeng, Amy Xu, Adam Zhang, Chen Xie, Yiwei Si, Runyi Yao, Fan Yang, Adrianne Kwon,

and Ruhong Jin. I will forever miss our food parties. Cooking for my Oxford gang was one of the happiest things for me. I want to thank Musashi Harukawa for those smoke-filled, trembling conversations we shared on so many sleepless nights outside our apartment. I am also very grateful for my current teams in Zurich and Geneva. As the youngest "baby" in our Zurich neurolinguistic lab, everyone takes care of me and makes me feel supported. My colleagues come from diverse academic backgrounds, and each of them warmly and patiently offers their expertise whenever I need help. Thank you, Enrico Varano, Natalia Bekemeier, Huw Swanborough, Alejandra Hüsser, and Christopher Ritter. I am deeply grateful for our Geneva Dynoball lab too. Thank you, Martina Dordijevic, Victor Ferat, Sophie Slaats and Maya Thille for your support, celebrations and gifts when I finished my Viva. We are the hive mind!

It's never easy to maintain a long-term friendship with someone as lazy as me, so I would like to give my genuine thanks to my friends in China who make their effort to keep me in check. I would like to say thank you to Yaolin Li and Qianru Guo for always being there, sharing our exciting news as well as awful ones. I would also like to thank my childhood friends Fangfang Liu, Yuanyuan Liu, Ziwen Liu, Mingxuan Liu, and Yan Zheng for supporting me through ups and downs for the last twenty years. You are my source of power, no matter you are near or far away from me.

My special thanks go to Jiayi Wei. She is my person. Even with my excellent autobiographical memory, I cannot recall exactly how many wonderful days and big moments we have had with each other. I'd thank her for many things: for cheering me up, for calming me down, and for watching me go crazy but still understanding me, but most of all, I want to thank her for always being with me in my wildest dreams, asleep and

awake. It makes me fearless to take risks and try new things in my life, knowing someone always sees me and believes in me. This is the luckiest thing that has ever happened to me.

It feels incredible to look back and see how I have been supported and loved. Life changes dramatically from moment to moment, but everything I mention in this acknowledgement has, somehow miraculously, been there for me all along. As I move on to the next chapter of my life, I hope to carry this passion, guidance, support, and love with me wherever I go.

Probably some final notes, thanks to X Japan, Joey Yung, Denise Ho, Snow Patrol, Collage, Florence + The Machine, and Arctic Monkeys for soundtracking my days and nights while I was writing this thesis. Their music kept me company through the longest hours and reminded me that, beyond the pages of my work, there is always rhythm, emotions, and life waiting for me.

20 November 2025,

On the first snowy day in Zurich

ABSTRACT

This thesis investigates the capacity of predictive mechanisms in real-time language comprehension to use subtle linguistic cues. Language comprehension, inherently sequential, involves incremental interpretation of linguistic inputs. However, predictive processing allows comprehenders to anticipate upcoming linguistic structures beyond the linear order, offering potential cognitive efficiency gains, particularly in long-distance dependency constructions. This study aims to understand the extent to which predictive strategies are employed, the level of detail comprehenders can predict, and the condition under which prediction occurs, through both behavioral and neurobiological measurements in Mandarin Chinese and English.

This thesis comprises three independent yet interrelated studies focusing on different linguistic cues that might trigger predictions. Study 1 examines whether Chinese classifiers, which constrain animacy without specifying particular lexical items, elicit semantic feature predictions. Using Representational Similarity Analysis on EEG recordings, Study 1 provides neurobiological evidence that comprehenders can predict abstract semantic features beyond specific lexical items. Study 2 extends these findings by exploring whether animacy-constraining classifiers can guide comprehenders to predict structural elements, i.e., gap sites in head-final relative clauses in Mandarin Chinese. Both eye-tracking and self-paced reading results demonstrate the comprehenders'ability to utilize classifiers to modulate active gap search in the absence of head noun fillers. Study 3 examines the effect of a less understood element, presuppositional constraints, on modulating active gap search. It compares negative island constraint, a presuppositional constraint on filler-gap dependency formation with strong complex NP island constraint, however, the results suggest that negative island constraint cannot be rapidly used by the real-time parser to inhibit active gap filling in the same way strong islands do.

Results across the three studies suggest that predictive structure building in real-time language processing is dynamically modulated by subtle linguistic cues, and cues at different levels might

take their effects with different timings: syntactic and semantic cues have more immediate effects while pragmatic cues take longer to compute. These findings contribute to a nuanced understanding of real-time language comprehension, teasing apart various factors affecting predictive structural building in dependency constructions.

LIST OF ARTICLES

Article one:

Huang, Z., Feng, C., & Qu, Q. (2023). Predicting coarse-grained semantic features in language comprehension: evidence from ERP representational similarity analysis and Chinese classifier. *Cerebral Cortex*, *33*(13), 8312-8320.

Article two:

Huang, Z. & Husband. E.M. (in prep). Classifier as a cue for structural prediction in processing Chinese head-final relative clauses.

Article three:

Huang, Z. & Husband. E.M. (in prep). Negative islands do not block active gap filling.

TABLE OF CONTENTS

Cover article

Cover Article

1 Introduction

Real-time language comprehension for humans is inherently sequential, reflecting the way we perceive linguistic information from the external world. Spoken language unfolds over time as speech sounds are produced syllable by syllable and these auditory inputs are processed continuously as they are received. Similarly, written words are also presented in a linear format, requiring sequential processing as our eyes move progressively across the text as our eye fixations have limited visual span. Therefore, it is fair to say that our perception of language in most forms is constrained by temporal order, where earlier inputs get processed earlier. It has been firmly established in the past several decades that comprehenders are always engaged in an incremental fashion to interpret linguistic inputs. To be more specific, comprehenders would like to interpret and integrate the currently available input immediately upon encountering it. Many psycholinguistic studies have provided robust evidence showing language comprehension is an incremental process (Altman & Kamide, 1999; Tanenhaus et al., 1995; Frazier & Clifton, 1986; Kamide et al., 2003; Staub & Clifton, 2006).

However, prediction allows us to jump out of this incremental linear order, generating pieces of linguistic information ahead of their actual presence, constructing the upcoming linguistic structure beyond the current scope, and thus facilitating the integration of future syntactic structures. One particular motivation for comprehenders to use such a predictive strategy is the need to process long-distance dependency constructions which frequently occur in most languages. Long-distance dependency normally involves a head and a dependent which are not linearly adjacent to each other, see (1) as an example. To correctly comprehend the sentence, the comprehenders need to realize that the wh-phrase “what” is analyzed later as the direct object in the clause “his kids like __”. In this case, the fronted “what” is called the head and its canonical grammatical position as the direct object is the dependent. In English, the canonical position is not overtly marked and

is represented by the blank space because it is displaced. However, despite the absence of an explicit direct object at the canonical position, comprehenders must infer this dependency to correctly interpret the sentence.

(1) The father knows what his kids like __ for dinner.

The comprehension of this non-linear association between the wh-phrase and the gap poses a challenge for a pure bottom-up processing strategy to achieve full incrementality, because when encountering the first element, normally the head, the integration of this element does not only involve previously built structure, but also involves upcoming structure. A strictly bottom-up parser, which focuses on the integration of the current element with the previous context, would process each element in isolation, delaying the establishment of the dependency until the gap is encountered, because this referential relation has nowhere to be projected onto the previous structure upon encountering the first element (normally the head). In this way, the referential integration of the first element has to be maintained until the comprehender parses the gap site and then be retrieved and revised from the working memory. An alternative, more efficient approach to deal with the non-linear referential relations involves active prediction: comprehenders anticipate the presence of a gap and immediately establish a provisional dependency upon encountering the wh-phrase. The realization of the latter possibility requires a certain level of active predictive mechanism as it requires reasoning about linguistic elements that have not yet appeared. Many psycho-/neuro-linguistic studies have provided robust experimental results supporting the predictive hypothesis, that, instead of waiting for the dependent/latter corresponding part to show up, comprehenders would initiate the structural building of this dependency right upon processing the head/first part and actively search for the corresponding part (e.g., Fodor, 1978; Crain & Fodor, 1985; Traxler & Pickering, 1996; Wagers and Phillips, 2009).

In fact, more and more results from recent psycholinguistic studies demonstrate the involvement of predictive processing in language comprehension at various linguistic levels including phonological, orthographical, syntactic and semantic representations. Incremental integration of current input and expectations about the upcoming input together make real-time language comprehension a mixture of both bottom-up and top-down processing, and therefore more efficient. Some language processing models even consider prediction to play a fundamental role in language processing (e.g., Pickering and Garrod, 2013; Pickering and Gambi, 2018; Clark, 2013). Nonetheless, critical questions remain: to what extent do comprehenders engage in predictive processing? How detailed are the predictions they generate, and under what circumstances are they willing to predict? To what extent does grammatical knowledge play a role in prediction decision-making? A large number of studies which support predictive processing in language comprehension rely heavily on the context in the experimental paradigm which elicits prediction of specific lexical items (a review of the predication studies: Kuperberg & Jaeger, 2016). The highly predictable context leading to specific lexical items (e.g., In the morning I prefer to have a cup of coffee with sugar and __) is very rare to be found in natural language and therefore it is unclear whether prediction can be generalized as a common strategy for the processing of natural language materials (Luke and Christianson, 2016). It also leaves open whether prediction at various abstract linguistic levels is independent from the prediction of concrete lexical items and how grammatical knowledge plays a role.

Structural prediction, generally meaning that the comprehenders can project or pre-build pieces of sentence structures ahead of the appearance of their lexical forms, provides a perfect test ground for investigating the relationship between grammatical knowledge and the cognitive mechanisms for prediction. In the theoretical linguistic framework, sentence/language structure is considered as a nested and layered hierarchical construction, where words are grouped into larger constituents (normally phrases). If we take into account the dependency relation as illustrated in the example (1), the syntactic

structure is a non-local cross-hierarchical construction. Thus, the processing of bottom-up lexical input may cue and activate higher-order constituents, prompting the anticipatory construction of syntactic structure that extends beyond the immediately available input. Since the linguistic cues might be subtle and the context is often less constraining for syntactic structure, such predictions about sentence structure can be risky and costly. Do comprehenders cast a top-down structure of the whole syntactic constituent upon the first bottom-up input of the constituent, regardless of risk? Or are they strictly bottom-up, focusing mainly on the current input integration with the existing syntactic constituents? Or is there a balance between the two extremes that only certain cues can trigger the comprehenders to make more predictions? And if so, how detailed can the predictions be? In short, the general question of interest for this thesis is, what are the limits for predictive mechanisms underlying real-time sentence structural processing? Three experimental studies have been conducted to utilize various subtle linguistic cues or constraints for generalizing structural predictions about the upcoming language information, and see if comprehenders are able to or willing to make efforts to generate more detailed predictions based on these cues.

Before we dive into the empirical studies, this thesis will first review the discussion about prediction in language comprehension in the literature, focusing on three key aspects: i. the triggers and timings of prediction (i.e., when is the prediction generated?), ii. the content of prediction (i.e., what is predicted?), and iii. the mechanisms for prediction (i.e., how is prediction implemented?). Then this thesis will discuss how the experimental work in this DPhil study can contribute to the current exploration of language prediction and present an overview of each individual experimental study.

1.1 Theoretical Perspectives on Predictive Processing

Prediction has increasingly been recognized as a crucial mechanism in real-time language comprehension. However, the term 'prediction' may have different meanings in

different contexts. Here in this thesis, we consider prediction as specific computation operations to actively generate linguistic representations about upcoming words or structure before the actual input is encountered. Thus the first keyword for prediction is “active”. Whether the processing decision or operation is active differentiates “prediction” from another closely related notion “integration”. Many facilitatory or inhibitory effects observed in the existing empirical studies can be, on one hand, interpreted as successful or unsuccessful predictions generated at an earlier point, or, on the other hand, can be interpreted as the current input requiring fewer or more resources to integrate itself into the existing sentence structure. Kazanina (2017) addressed the distinction between prediction and integration by examining whether genitive object triggers a prediction for a negated verb in Russian or whether a negated verb by itself is easier to be integrated into the context. Her experiment showed that in the object-first configuration, genitive case on the preverbal object leads to a clear facilitatory effect at the negated verb, which is consistent with active prediction of negation and the verb based on case marking. While in the negation-first configuration, where the same genitive–negation association exists but prediction is not needed to maintain a connected structure, there is no corresponding facilitation at the object. These results demonstrated that integration accounts cannot explain why facilitation appears only when prediction is structurally necessary for incremental parsing. Based on it, Kazanina also pointed out that the core motivation for predictive processing is to achieve full incrementality, meaning that the predictive mechanisms “aim to attain as fully incremental a structure-building process as possible while ensuring grammatical licensing”. Crocker (2002) also explicitly characterizes the human sentence parser as “even more eager than” a strictly incremental integrator, “constructing anticipatory hypotheses about what is likely to follow”. Building on these, predictive processing goes beyond reactive integration. It is a strategy that enables the sentence processor to maintain a fully connected and incrementally updated structural representation of the sentence, even before bottom-up confirmation arrives. In this sense, predictive mechanisms serve as a solution to the challenge of incrementality. To outline a

predictive processor, the critical questions to be asked are not what happens at the current input, but what happens before the current input: What are potential triggers for the prediction? What are the representations generated by prediction? What is the appropriate algorithm to implement a predictive mechanism? There are enormous theories and models proposed to address these questions, each providing an effective solution to account for specific linguistic phenomena. In the following parts of this session, we will review these aspects perspectively.

1.1.1 When is the prediction generated: the timings and triggers of prediction

The timing of prediction is closely related to the triggers of prediction: is the predictive process triggered or cued by specific linguistic items, or is it continuously active? This question has been approached from two broad perspectives: the cue-based view, which sees prediction as triggered by specific linguistic input, and the continuous view, which treats prediction as an always-active, dynamically modulated process. These views are not mutually exclusive, and empirical evidence suggests that both types of prediction may operate, potentially at different levels of representation.

The dominant view in much of the psycholinguistic literature is that linguistic cues serve as discrete triggers for prediction. For instance, as discussed earlier, Kazanina (2017) shows that genitive case marking on a preverbal object in Russian can trigger the prediction of a negated verb, but only when this prediction is structurally required for maintaining an incremental parse. In configurations where the negation marker precedes the object, and thus prediction is not necessary, no facilitation is observed. Another example of a similar cue-triggered effect was observed in Crocker (2002), where case marking in German was shown to prompt anticipatory eye movements towards expected arguments. These predictive behaviors emerge only after relevant linguistic cues become available, reinforcing the view that the time course of prediction effect is tightly locked with informative input. On the other hand, probabilistic models of language comprehension,

such as those developed by Roark (2001), Jurafsky (1996) and Levy (2008), treat prediction as a continuously active process. Under this continuous prediction view, at any given moment, the sentence parser stays context sensitive and maintains a distribution over possible upcoming words or structures, weighted by prior knowledge and contextual constraints. One example of such continuous prediction is Roark's top-down probabilistic parser (2001). It assigns probabilities to predicted syntactic expansions, continuously updating these distributions with each new word. Furthermore, studies using surprisal-based metrics (e.g., Smith & Levy, 2013) argue that facilitation and processing difficulty can be predicted from continuous gradient expectations. Levy's expectation-based model adopts a maximally continuous view of prediction: prediction is no longer an optional strategy that the parser engages in at particular points, but the default mode of processing, in which comprehenders continuously maintain probabilistic expectations over upcoming words and structures. Within this framework, surprisal serves as the central linking hypothesis between expectation and processing difficulty: the more unexpected a word is given the current context, the greater the processing cost it incurs. In this sense, surprisal replaces discrete, grammar-based prediction cues with a gradient, probabilistic metric that is defined at every input position. Empirically, expectation-based models receive support from a wide range of studies. Surprisal estimates derived from probabilistic grammars and language models reliably predict reading times and eye-tracking measures in self-paced and naturalistic reading (Hale 2001, 2003; Levy 2008; Demberg & Keller 2008; Frank et al. 2013; Smith & Levy 2013). At the same time, many classic syntactic phenomena, including garden-path effects and relative clause asymmetries, can be re-analyzed as consequences of low-probability continuations under an expectation-based parser (Levy 2008; Levy et al. 2013). Together, these findings support the core claim that human sentence processing is tightly guided by graded probabilistic expectations. However, rather than viewing the cue-triggered prediction and continuous prediction as competing accounts, many frameworks (e.g., Crocker & Brants, 2000; Levy, 2008) treat them as complementary. Cue-triggered prediction may reflect rapid structural commitments made

when certain linguistic cues demand it (e.g., case-marked NPs, wh-elements), while continuous prediction may reflect gradient expectations maintained over the course of comprehension, modulated by lexical frequency, context, and world knowledge.

Whether prediction is cue-triggered or continuous based on surprisal metrics, or with a dual nature, a further question remains: what kind of information is allowed to serve as a cue or to sharpen and reshape the expectation distribution? Do certain cues carry greater weight and enjoy privileged status in guiding predictive decisions? In the discussion of cue weight, syntactic cues are often considered to be the more privileged, even deterministic, compared with other cues. This syntax-first view naturally aligns with the cue-triggered prediction view, with a narrow sense that, the cues can license predictive operations are typically restricted to syntactic information, such as case marking, grammatical agreements, and phrase structure rules, because these are seen as the most reliable indicators of forthcoming structure, and thus can demand prediction for the needs of structural incrementality. Garden path models by Frazier (1987) and the parsing principles proposed by Kimball (1973) are under this “syntax first” view. The garden path effects at the disambiguating regions and the filled-gap effects (Stowe, 1986) suggest that the initial analysis of the parser is rather syntactic-heuristic. Similarly, Crocker (1994) believed that at least the “first pass” parsing should be syntactic autonomous, as the semantic interpretations of the inputs do not occur soon enough to influence the parser’s initial decisions. However, strict syntactic autonomy has been increasingly challenged by empirical evidence showing that non-syntactic cues can influence processing even before disambiguation occurs. Findings from lexical bias, plausibility manipulation, and visual world studies all point to a more interactive system in which prediction is not solely driven by grammar but can be initiated or shaped by a broader range of cues. For example, Altmann & Steedman (1988) show that thematic fit and semantic plausibility can immediately bias structure: when in a prepositional phrase (PP) attachment ambiguity situation (e.g., The burglar blew open the safe with…), the semantic meaning of the noun

((e.g., … with the dynamite vs. … with the new lock) can guide the attachment preference before a purely syntactic preference is settled. Tanenhaus, Carlson, & Trueswell (1989) showed that lexical argument-structure biases and thematic constraints can influence the earliest resolution of NP/S ambiguities (e.g., The witness admitted/confessed the mistake …): verbs that strongly prefer a sentential complement (e.g., admit) drive readers toward a CP continuation, whereas NP-biased verbs (e.g., confess) push an NP analysis; critically, these effects appear at or before the disambiguating region in eye-movement measures, indicating that the parser utilizes verb-specific semantics and subcategorization as it commits to a structure. Together, these findings motivate constraint-based frameworks in which the chosen parse is the best-supported option given all available cues, rather than the output of a strictly syntax-first stage followed by later semantic repairs. Lexicalist Constraint-Based Parsing by MacDonald, Pearlmutter & Seidenberg (1994), Probabilistic Constraint Integration by Jurafsky (1996), Verb Subcategorization Constraint Models by Trueswell, Tanenhaus & Garnsey (1994) and many other models testing different cues all contribute to the constraint-based parsing framework. In this framework, the parser computes the most probable interpretation at each point by weighing competing constraints, often in real time. For instance, in a sentence like “The reporter interviewed the daughter of the colonel who...”, syntactic structure might favor attaching “who” to “the colonel”, while semantic plausibility or discourse prominence might suggest “the daughter” is the intended antecedent. In such cases, the parser evaluates both structural configurations simultaneously, and the final commitment depends on which interpretation is better supported across all levels of representation. This constraint-based conflict resolution is often formalized in probabilistic models, such as Jurafsky’s (1996) Bayesian parser or Levy’s (2008) expectation-based framework, where the parser computes the likelihood of different syntactic structures given the input and prior experience. The parser is seen as a rational statistical learner, dynamically integrating cues to minimize surprisal and maximize interpretive coherence. In short, the interactive parser resolves cue conflicts not by adhering to a fixed hierarchy

of cue priority, but by evaluating how much support each cue lends to a given interpretation, dynamically adapting to both the immediate linguistic input and broader contextual expectations. This leads to graded processing effects, such as temporary ambiguity, processing slowdowns, or even garden-paths, when the competition is close or when a strongly supported structure later turns out to be incorrect.

1.1.2 What is predicted: the content of prediction

Another defining aspect of language prediction is what linguistic representations can be generated. The content of expectations can range from specific lexical predictions to syntactic categories, and also to abstract semantic feature bundles. Different theories differ in the grain and nature of those expectations. Early on, Kimball (1973; 1975) suggested that readers may not generate expectations for exact words but can predict the syntactic categories, i.e., part of speech or phrase type of the next element in a sentence based on the context. For example, upon reading “Joey devoured…” one confidently expects a noun phrase as the direct object to follow, whereas after “John devoured the very…” one expects an adjective and a noun to appear next, since “the very” signals a noun is being modified. Similarly, a complex sentence segment like “the boy who the king…” licenses the expectation that two verb phrases will eventually appear, one for the relative clause “who the king (…verb)” and one for the main clause “the boy (…verb)”. even if additional words might intervene before those verbs materialize. In short, the category and structural role of upcoming constituents from different layers of the hierarchical structure can be anticipated. Many models of predictive parsing explicitly build such expectations into the parsing process. For instance, Crocker’s framework (2002) holds that the parser can “enter nodes into the representation of the sentence’s syntactic structure…before encountering the corresponding lexical input”. The nodes typically refer the phrasal constituents or placeholders required by the grammar. Schneider (1999) likewise allows the parser to posit unseen syntactic heads based on features of already-seen words. In his model, encountering a noun with a certain case marking (e.g. a dative-

case noun in a head-final language) leads the parser to expect some governing head will follow to license that case. Crucially, the prediction might be underspecified – meaning that the parser only knows a dative-case licenser is needed, but might not further predict the syntactic category of the licenser. In all the accounts mentioned above, the discussions about predicted elements are pure syntactic in nature (phrasal nodes, clausal nodes, or syntactic heads with certain features). However, beyond syntactic prediction, abstract features that are frequently associated with certain syntactic categories might also be predicted. For example, an ERP study by Wicha and colleagues (2004) showed early costs when the gender mark of the article mismatched the gender mark of the predicted noun in Spanish, demonstrating a feature-level pre-activation during language comprehension. Szewczyk & Schriefers (2013) used Polish sentences where pronominals carry morphological animacy features. Items whose animacy marking conflicted with the predicted noun's animacy produced ERP costs at the pronominals, demonstrating prediction of animacy features ahead of the noun. McRae, Spivey-Knowlton & Tanenhaus (1998) combined experimental and modeling approaches to show that readers immediately use event knowledge/thematic roles fitting (e.g., which entities are likely to be Agents or Patients for a verb) during ambiguity resolution, biasing structure and role assignment before disambiguating input, suggesting thematic role-level predictions. Furthermore, DeLong, Urbach & Kutas (2005) used English a/an articles and showed that article-elicited N400s scaled with cloze for the upcoming noun's initial sound, implying graded pre-activation of phonological form prior to the noun itself. Kuperberg & Jaeger's review (2015) highlights that comprehenders predict probabilistically and across levels, from high-level semantics down to lower-level lexical-form features. However, it still remains unclear how such fine-grained feature prediction can be elicited. Most of these studies created a context that makes certain lexical items highly predictable. It is yet to be explored, when given some weakly-constrained context, what content, and how much fine-grained feature prediction can happen.

1.1.3 How is prediction implemented: the mechanisms of prediction

Predictive processing involves the generation of representations prior to receiving the corresponding bottom-up input. The key issue is how to derive these representations while preserving strong incrementality and flexible structure building on the algorithmic level. Numerous models of human language processing have been proposed, adopting different strategies for syntactic projection and consequently differing in the extent to which they permit prediction. As discussed earlier, the core motivation for predictive processing is to achieve immediate interpretation and integration during real-time language comprehension. To achieve immediate incrementality in more complex language constructions, such as head-final constructions, a satisfactory architecture for an incremental human sentence parser must be at least partially top-down, projecting upcoming syntactic structures based on available inputs (Crocker, 1994). Therefore, head-driven models, as proposed by Abney (1989) and Pritchett (1992), struggle to deliver fully incremental processing for head-final constructions, as they assume structure integration and attachment has to be initiated by the phrasal heads, resulting in the parser has to give up immediate incrementality in sentences where crucial heads appear late. Top-down parsers, on the other hand, can support incremental processing in both head-initial and head-final constructions. In a top-down parser, syntactic processing begins from the highest-level category and proceeds downward, projecting expected constituents before the relevant input is encountered. This architecture embodies a strong form of prediction: the parser commits early to structural expectations derived from grammar. The Garden-Path model (Frazier & Fodor, 1978; Frazier & Rayner, 1982; Rayner et al., 1983) holds a classic top-down view: it initiates a single analysis guided by Minimal Attachment and Late Closure and reanalyzes the structure when the bottom-up input contradicts these early commitments. Beyond the heuristic garden-path model, probabilistic top-down models such as Roark (2001) maintain a connected top-down derivation while ranking multiple partial parses in a beam, thereby preserving strong pre-head projection but with softer commitments and explicit next-word probabilities; similarly, principle-based top-

down approaches (e.g., Crocker, 1994; 2002) predict the functional heads and posit gaps early, providing a mechanistic account of pre-head structure building in languages where arguments often precede their verbal heads. However, top-down parsing often requires excessively strong, often error-prone commitments, especially in structurally ambiguous and head-final configurations. Complementing top-down approaches, a left-corner parser integrates bottom-up and top-down processes. It initiates structure building when the "left corner" of a phrase is recognized in the input. From that point, it predicts the higher-level structure that could dominate this constituent while continuing to process the bottom-up input. This hybrid mechanism maintains incrementality but grounds prediction in the actual input, thus avoiding excessive or premature commitments typical of fully top-down systems (Rosenkrantz & Lewis, 1970; Johnson-Laird, 1983; Crocker, 1999). Crocker (1999) develops a left-corner architecture for human parsing that uses recognized left corners to maintain a connected, incrementally interpretable structure and to explain locality effects while avoiding global reanalysis. Extending the left-corner model family to a lexicalized formalism, Demberg et al. (2013)'s PLTAG model introduces Prediction Trees in a TAG framework: treelets anchored by observed words can license predictions of higher structure that are later verified by incoming input, achieving a left-corner–style balance between early projection and evidence-driven confirmation. This model posits several different prediction trees in parallel, and a separate verification system confirms the correct predictions and converts them into lexically licensed ones. Although left-corner parsing integrates both bottom-up and top-down approaches, it is not problem-free: standard left-corner parsing runs into trouble when facing massive left-edge ambiguity. An initial NP sequence is compatible with many continuations, and thus the left-corner parser is forced to pick one category-level structure, e.g., a simple clause structure, and then frequently reanalyze it into a complement, adjunct, or relative clause, even though such continuations are not in fact experienced as especially difficult by speakers.

To account for the left-edge ambiguity problem discussed above, Schneider (1999)

proposes a variant of left-corner parsers that has sufficient flexibility for prediction commitment. This parsing algorithm is called SPARSE. Instead of projecting full phrasal categories like NP or VP as soon as a left corner is seen, SPARSE operates over bundles of syntactic and semantic features, such as category, case, number, animacy, and licensing features that specify what a head can license or must be licensed by. Words enter the derivation as these feature bundles, and structure is built only when licensing relations between features can be satisfied. When the feature licensing is not satisfied, to keep the structure connected, the parser may posit an underspecified predicted head characterized only by the minimal features needed (e.g. “a head that licenses nominative case”), rather than committing to a specific category or lexical item. In this way, SPARSE remains fully incremental and can maintain a single connected structure. Since its predictions are feature-based and underspecified, it keeps multiple downstream continuations available and dramatically reduces the need for large-scale reanalysis in the densely ambiguous, head-final strings that pose a serious challenge to standard left-corner parsers.

Among the parsing models that allow a decent amount of prediction during incremental language processing, they diverge from several core axes of language prediction as we discussed in the previous sections — what information licenses prediction, what the parser is able to predict, and how strongly the parser commits. We can very broadly situate them along the three core axes of prediction. Garden-Path parsers are classical top-down parsing and are paradigmatically syntax-first: prediction is licensed almost exclusively by grammatical category and phrase-structure principles, and the content of prediction is largely a structural skeleton, projecting major phrasal nodes and attachment sites rather than rich semantic detail, and the commitments are hard, in the sense that a single analysis is pursued until contradictory bottom-up input forces costly reanalysis. Principle-based top-down models such as Crocker’s retain a syntax-first flavor but allow somewhat richer structural content, in the sense that functional heads and gap positions

are projected in advance, while still committing relatively strongly compared to probabilistic approaches. Probabilistic top-down parsers like Roark's are still primarily syntax-driven in their cues but often incorporate lexical and frequency information. They support more detailed prediction content, including next-word and category probabilities, and as a consequence of maintaining several ranked analyses in parallel, they do not normally commit to one structure. Expectation-based surprisal models, mostly also probabilistic top-down like Hale (2001) and Levy (2008), while a few are left-corner like Boston et al. (2011), are more interactive in terms of prediction cues compared with Roark's probabilistic parser. Standard left-corner parsers and lexicalized variants such as PLTAG move further toward interactive cues, because prediction is tightly constrained by what has actually been seen in the input as well as grammar. Their prediction content is somewhere between a skeleton and a detailed structure, such as partial spines or prediction trees that anticipate upcoming heads and attachment sites. Their structure commitments are softer than in fully top-down systems because multiple predicted continuations can be maintained and later verified. Finally, SPARSE, with its feature bundles and underspecified predicted heads, is essentially principle-based and grammatical driven in the sense of separate serial syntax/semantics modules talking to each other, but it is de facto interactive as they incorporate the "non-syntactic" information including semantic features and sortal information (modified SPARSE model proposed by Yoshida, 2006), into the grammatical cues that trigger prediction. SPARSE models predict underspecified, feature-level structures rather than fully fleshed-out phrase markers, and they embody a very soft commitment strategy in which multiple continuations remain open until licensing relations force a more specific structural choice. Constraint-based and other strongly interactive models push this logic even further, treating syntactic, semantic, pragmatic, and probabilistic information as jointly licensing highly detailed predictions while keeping commitments gradient and revisable throughout processing.

1.1.4 The theoretical questions of this thesis

This thesis aims to test the limits of prediction processing with three experimental studies and provides empirical evidence for outlining or fine-tuning the predictive mechanisms for human sentence processing according to the core aspects of language prediction as discussed above. This thesis explores the questions of what information licenses prediction, what the parser can predict, and how strongly the parser commits when with very limited or subtle linguistic context. These are the core issues that define a predictive language processing model for human sentence processing. Our general idea is to utilize various subtle linguistic cues or grammatical constraints that require linguistic awareness and grammatical reasoning to generalize predictions about the upcoming language information, and see if comprehenders are able to or willing to make efforts to generate more detailed predictions based on these cues. In addition, this thesis also explores the interaction of linguistic cues from different levels: when multiple cues are available for making syntactic predictions, how the parser resolve conflicts among these cues? Furthermore, by probing into how detailed the real-time prediction can be, we might gain insights into the timing of predictive processing in real-time language comprehension. The earlier a prediction is made, the more details it can potentially include. To this end, three studies have been conducted. The studies are independent and focus on different linguistic cues and their effects on predictive structural building. However, these three studies all center around the general question of the limits of prediction in real-time language processing, and they all observe the prediction effects on the dependency relations. Article 1 mainly addresses the question of prediction content. Article 2 also addresses the granularity of prediction content, as well as prediction triggers. Article 3 puts more emphasis on what cues trigger prediction, especially when it comes to the interaction of cues from different levels of linguistic representations. All three studies, with high-resolution experimental measurements, i.e., EEG and eye trackers, also provide opportunities to discuss the question of when language predictions happen.

1.2 Overview of the experimental studies in this thesis

Several questions of general interest of these studies should be answered in this introduction, before diving into a detailed introduction of each study. The general questions include: why focus on the effects on the dependency relation? What linguistic cues are examined in each study? And how are the three studies related to and/or different from each other?

The first question to be answered is why the studies choose to observe the prediction effects on dependency relations. The processing of dependency relations is a very good start to discuss prediction-related questions because:

a. The realization of the immediate integration of a dependency construction in real-time processing indeed involves active prediction as noted earlier. The full interpretation of the first element in the dependency relation relies on the not-yet-appearing second element. Thus prediction is required to construct such a non-linear structural integration.
b. The non-linear nature of dependency relations (two parts which are not linearly adjacent to each other) allows us to assess not only whether but also when and how far in advance predictions can be made.
c. The prediction of dependency relation is essentially a structure prediction which does not require the prediction of specific lexical items. For example in (1), the referential dependency between the wh-phrase and the gap site needs to be established but the referent does not have to be specified.
d. Dependencies might involve various cues at different linguistic levels that might trigger predictions about upcoming information, allowing us to understand the role of specific cues in guiding predictions and to identify the conditions under which predictive mechanisms are activated, which contributes to a broader understanding of how selective or generalized the predictive mechanisms are in different linguistic contexts.

Second question: what linguistic cues are examined in each study? Study 1 makes use of Chinese classifiers which provide animacy constraints on the nouns that they modify (for a detailed explanation see section 2.2.1). The animacy-constraining classifiers do not lead to the predictions of specific lexical items because each classifier can modify multiple different nouns, but these classifiers can indicate certain shared semantic features among the nouns that they modify. For example, all the nouns that the classifier “wei” can modify are all human nouns like “laoshi” (“teacher”), “yisheng” (“doctor”) etc. If comprehenders can utilize this animacy-constraining cue provided by the classifier, they might be able to predict some of the semantic features of the upcoming nouns even without knowing the specific nouns. Study 2 builds on the animacy constraining classifiers and further utilizes the mismatch cue between the classifier and its following word (for a detailed explanation see section 3.2.2). The mismatch implicitly suggests that the classifier is dislocated from the noun that it modifies and there is a subordinate clause inserted in between. If comprehenders can recognize this cue in real-time processing and use it in combination with the animacy constraint, they might be able to predict different types of relative clauses. Study 3 uses negative island, a pragmatic presuppositional constraint on the extractions out of island domains (detailed explanation see section 4.2.2), to see if comprehenders can avoid “over-prediction” in the island domains. Negative island effects are defined and observed mainly in off-line (opposite to the online and real-time processing) grammatical judgment tasks, therefore it remains unknown if comprehenders can modulate their eager predictions based on pragmatic cues. More detailed introductions about these prediction cues will be given in later sessions. These linguistic cues do not heavily depend on high co-occurrence probabilities but require immediate linguistic awareness and reasoning based on grammatical rules to elicit predictions, thus, they can help us understand better whether grammatical knowledge can guide real-time language processing decisions.

Then the third question that might be of general interest is, how are the three studies related to and/or different from each other. The linguistic cues for Study 1 and Study 2 are

more closely related as Study 1 establishes the foundation for the potential effects of the linguistic cue in Study 2. Study 1 seeks direct neurobiological evidence for the animacy prediction effects in the classifier-noun phrase dependency. Then building upon the semantic dependency relation between a classifier and its head noun, Study 2 adds filler-gap dependency processing into the picture and further explores whether the semantic feature prediction triggered by a classifier can be used to guide the comprehenders to predict gap sites in a relative clause construction in Mandarin Chinese. Study 3 also targets at the processing of filler-gap dependency as in Study 2, but it differs in both the type of cues and the language groups, focusing on English and investigating whether pragmatic presuppositional constraints, such as negative islands, can modulate predictive behavior to prevent over-prediction in filler-gap dependencies. It provides a different perspective for us to explore whether the level of representations of the linguistic cues matters when it comes to real-time predictions. Pragmatic cues are generally considered to be highly context-dependent and more complex to compute for a real-time parser during language comprehension. By exploring various cues, we can learn more about how differently the linguistic cues from various levels are handled in real-time.

To summarize, the three studies in this thesis focus on the structural building of long-distance dependencies to explore what cues can be used to trigger active prediction for structure building and how detailed or what aspect of information can be predicted during incremental language comprehension. Three individual experimental studies bring evidence of predictive processing from different perspectives, including different kinds of dependencies (phrasal dependency, clausal filler-gap dependency and prepositional island constraint to clausal filler-gap dependency), different languages (Mandarin Chinese and English) and also different measurement techniques (for behavioral measurements including grammatical judgment tasks, self-paced reading tasks and eye-tracking, to neurobiological measurements like Electroencephalography (EEG)).

In the following chapters, this thesis explains the research questions and rationales for each experimental project respectively, and then the research background will then be introduced. A summary is also presented in each chapter to give an overview of the study. Then an integrative discussion will be given in this cover article, including the main findings of each study, a general discussion from an overall perspective (detailed discussion please see in each article), limitations, and future directions. The three individual studies are written in the format of manuscripts for publication and are attached as appendixes to this cover article.

2 Article 1: Predicting Coarse-grained Semantic Features in Language Comprehension: Evidence from ERP Representational Similarity Analysis and Chinese Classifier

2.1 Research question: how abstract can prediction be?

As outlined in the introduction, the central objective of this thesis is to investigate the limitations of predictive mechanisms in real-time language comprehension. Specifically, this article examines how abstract predictions can be, independent of specific lexical predictions. A large amount of research in language prediction relies heavily on the prediction of specific lexical items (e.g., Wicha et al. 2004; Ito et al. 2018; Li et al. 2022). They often use highly constraining context cues or semantically interrelated lexical items to narrow down the predictions to one or one set of "high-cloze" lexical items. The results of these studies show that when specific lexical concepts are elicited, the prediction outcomes can be rather fine-grained, including semantic features (Altmann and Kamide 1999; Federmeier and Kutas 1999; Lau et al. 2013; Wang et al. 2018, 2020), phonological features (DeLong et al. 2005; Vissers et al. 2006; Li et al. 2022), written form (Laszlo and Federmeier 2009; Kim and Lai 2012), and morphosyntactic features (Van Berkum et al. 2005; Dikker et al. 2009, 2010). Then one question that comes naturally about these observations is: is pre-activation able to target features independently of any particular lexical item? Or are these abstract linguistic feature representations pre-activated only when a specific lexical item is pre-activated? This question has received less attention in the research on language prediction. To disentangle the abstract feature prediction from lexical prediction, one possible solution is to use prediction triggers that do not lead to specific lexical prediction while also being reliable enough to elicit certain feature predictions. Some research has been done to investigate predictions beyond specific words. For example, Szewczyk and Schriefers (2013) conducted an ERP study to investigate whether comprehenders can predict broad semantically defined classes of words. Comprehenders are instructed to read short stories, and these stories were setting

up a context that was strongly biasing towards either an animate or an inanimate noun in the direct object position in the final sentence of the story. Event-related potentials (ERPs) were measured as participants read the adjectives that either matched or violated predicted grammatical features. The findings showed that violations of predicted features elicited an N400 component, suggesting that readers actively form predictions beyond individual word identity. A more recent work by Giskes and colleagues (2023) also explored the abstract prediction of morphosyntactic features during language processing by examining the comprehension of cataphoric pronouns in Dutch. They use eye-tracking to investigate whether readers generate morphosyntactic predictions (such as gender and number) based on cataphoric pronouns before encountering their antecedents. This study is also aiming at eliciting abstract feature predictions, but instead of using complex sentential context as cues, we try to set up minimum context by examining whether a single classifier can trigger feature predictions beyond words. Therefore, article 1 investigates whether Chinese classifiers can elicit semantic feature pre-activation without predicting specific lexical items. This study examines neurobiological evidence for semantic feature prediction by using EEG to measure the spatial and temporal brain activities.

The N400 component has been long considered to be one of the most important indices for language-related effects in EEG experiments. Some of the language prediction studies also target at observing the N400 component and interpret the reduced N400 amplitude as a sign of pre-activation of the target word. For example, when participants process sentences like "She spread the bread with butter," the word "butter" elicits a smaller N400 than a less expected word like "yogurt." This reduction is often taken as evidence that listeners or readers generate pre-activation of likely words, which then requires less neural effort to process when the prediction is confirmed (DeLong et al., 2005). However, the same N400 effects can also be interpreted through the lens of integration rather than prediction. Words that fit well with the preceding context are easier to integrate

semantically into the discourse. For instance, the word "butter" may elicit a reduced N400 not because it was explicitly pre-activated but because it aligns closely with the context, making integration seamless. Van Petten and Luka (2012) argue that the N400 reflects not just a prediction but also the ease of combining incoming information with prior linguistic and world knowledge, highlighting the overlap between these two processes. Since N400 effects take place after the presentation of the current input, to demonstrate prediction effects apart from integration effects, this study focuses on the analysis of the spatiotemporal patterns of brain activities prior to the presentation of the target words rather than the event-related responses that are inevitably affected by the current input.

Therefore, this study investigates whether Chinese classifiers elicit animacy feature predictions for their head nouns in classifier clauses using EEG recordings. The data analysis focuses on the brain activity patterns prior to the presentation of the head nouns using representation similarity analysis. In the following research background sections, we will focus on introducing a. why Chinese classifiers are used to elicit animacy feature predictions (section 2.2.1); b. why animate/inanimate distinction has been selected to be the semantic feature that we are interested in (section 2.2.2); c. and why representational similarity analysis has been used to detect the differences in the neuro activity patterns for animate and inanimate features (section 2.2.3).

2.2 Research background

2.2.1 Chinese classifier phrase and its animacy constraint

Mandarin Chinese is a numeral classifier language, which means that a classifier is obligatory when the noun phrase is introduced by a numeral, a demonstrative, or a quantifier (Li and Thompson, 1989). Most classifiers are encoded with physical or functional information that indicates the semantic properties, such as shape, size, or the status of their head nouns, thus each classifier can only modify a particular type of noun phrase when the semantic properties of the classifier match the classifier and the noun.

For example, the classifier “ming” in (2a) can only modify human noun phrases, especially respectful humans like teacher, doctor or the elder people, while the classifier “ben” in (2b) can only modify book-like noun phrases like book, magazine or album. It would be ungrammatical when the classifier and the head noun do not match with each other as shown in (3a) and (3b).

(2)

a. San ming lao shi
Three CL_{human} teacher
Three teachers

b. Zhe ben za zhi
This CL_{book} magazine
This magazine

(3)

a. * san ben lao shi
Three CL_{book} teacher
Three teachers

b. * zhe ming za zhi
This CL_{human} magazine
This magazine

The association between a classifier and a noun can be abstract or arbitrary to some extent. For example, a noun can be paired with different classifiers and a classifier can modify different nouns depending on how nouns are perceived in different contexts. What’s more, the classifier-noun association is sometimes robust while the semantic relation between them is obscure. However, a close semantic relation in classifier-noun pairs and the role of classifiers in predicting upcoming linguistic inputs has been attested in many studies using various paradigms. For example, comprehenders are presented with classifier-noun mismatch errors to investigate semantic integration processes during

Chinese sentence comprehension. A typical observation is the N400 effects on the nouns elicited by the classifier-noun mismatch, with larger negativity for mismatch trials reflecting the difficulty of integrating the lexical semantics into the representation at the higher level (e.g., Zhang et al., 2012; Zhou et al., 2010). The robust semantic agreement between a classifier and a noun, combined with ERPs, can provide an effective approach to examining the effects of prediction during language comprehension. In the research line of semantic prediction in Chinese, Kwon et al. (2017) investigated the role of classifiers in semantic predictions by manipulating whether a classifier embedded in a sentence matched or mismatched an upcoming expected noun. Kwon et al. observed the well-established N400 effect with enhanced N400 amplitude to unexpected nouns. More importantly, the N400 was also evident as early as the preceding classifier, suggesting the pre-activation of semantic features of nouns.

In this study, we select classifiers that are highly constraining for a semantic feature while low-constraining for specific lexical items to eliminate the lexical activation effects. To be more specific, we looked for classifiers that unambiguously modify animate entities, e.g., humans and animals, and that unambiguously modify inanimate entities, e.g., natural objects like stones and manufactured products like a telephone. By selecting these animacy-constraining classifiers, we can avoid the arbitrariness or obscurity between the classifier-noun pairs as mentioned earlier. We also made sure to exclude the classifiers that modify a narrowed group of entities. For example, the classifier “liang” can only modify vehicle nouns. So although it unambiguously leads to inanimate entities (vehicles), we still excluded it because it can also elicit specific lexical prediction.

2.2.2 Animate/inanimate distinction

The distinction between animate and inanimate entities is considered one of the fundamental categorical divisions in human cognition. Research in cognitive psychology suggests that this distinction arises from evolutionary adaptations that are essential for

survival. Humans tend to treat animate objects (such as animals and humans) differently from inanimate objects (such as tools or rocks), due to the potential for animates to be either threats or sources of social interaction. Studies in developmental psychology (e.g., Gelman & Opfer, 2002) show that children as young as one year old can distinguish between animate and inanimate objects based on motion patterns, intentionality, and biological processes. This innate sensitivity to animacy seems to be a building block for more complex cognitive processes, including language comprehension.

The animate/inanimate distinction affects how linguistic information is processed at various levels including lexical processing, syntactic structure building, and semantic and/or pragmatic interpretations. For lexical processing, research shows that words referring to animate objects (like "dog" or "person") are processed more quickly and occupy greater cognitive resources compared to inanimate nouns (like "chair" or "book"). A study by Caramazza and Shelton (1998) found that brain damage affecting specific categories (such as animals or tools) suggests that the brain processes animate and inanimate entities in distinct ways. Cree and McRae (2003) further suggest that these categories are uniquely represented in the brain due to their semantic properties, contributing to differences in processing speed and accuracy. The animacy of words can affect syntactic choices during sentence comprehension, especially the assignment of grammatical roles such as subject and object. Animate objects are often the subjects of sentences and are associated with actions or events that require volition or agency, whereas inanimate objects tend to be the objects or recipients of actions. Bock and Warren (1985) demonstrated that speakers tend to use animate subjects in their utterances, likely because animates are perceived as agents, capable of action, while inanimates are more passively involved. Moreover, Mak, Vonk, and Schriefers (2002) demonstrated that animacy effects when processing complex syntactic structures. In relative clauses, for example, people tend to find sentences easier to understand when the subject is animate and the object is inanimate (e.g., "The boy who chased the ball").

This suggests that animacy influences parsing strategies during sentence processing. Animacy affects pragmatic comprehension as well. Prat-Sala and Branigan (2000) showed that speakers tend to introduce animate entities early in a discourse and assign them prominent syntactic roles. In stories and conversations, animate entities often serve as discourse topics because they are seen as more salient and likely to participate in actions.

One important feature of the animate/inanimate distinction that is critical for this study rationale is, animate entities share more co-occurring semantic features than inanimate entities. Animate entities (e.g., “swimmer” and “lawyer”) share many semantic features, such as [can move], [can breathe], [can think], [can feel], etc. More shared features create a stronger intercorrelation within the animate category. In contrast, inanimate entities (e.g., “needle” and “water”) have their own distinct, unique features, resulting in less semantic overlap between them. This difference in feature clustering explains why the inanimate category includes a wider range of subordinate categories (e.g., furniture, vegetables, tools), whereas animate entities are more tightly grouped (McRae et al., 1997; 2005; Zannino et al., 2006). Computational models, including connectionist networks (Rogers & McClelland, 2008) and Bayesian approaches (Kemp & Tenenbaum, 2008) also suggest that differences in co-occurring features shape the taxonomic structure of animate and inanimate categories. These models demonstrate that animate entities, due to their higher semantic feature overlap, form tighter conceptual clusters. In contrast, inanimate categories—having more distinctive features—are organized more broadly, forming a larger number of subcategories. Dilkina et al. (2013), further confirms that semantic similarity between animate concepts facilitates faster and more accurate retrieval in both normal cognition and neural network models.

The semantic features are represented across widely distributed networks in the brain (Huth et al., 2016; Martin, 2016), thus the difference in semantic similarity between

animate and inanimate concepts should be reflected in their spatiotemporal patterns of neural activities. Differences in neural activity patterns are indeed observed which can explain category-specific impairments in patients with semantic dementia or other neuropathological conditions (Devlin et al., 1998; Tyler & Moss, 2001). Some studies such as those by Mahon & Caramazza (2011), confirm that the brain's representation of animate categories is more robust and interconnected. Thus the differences in neuron activity patterns regarding animate and inanimate concepts should be able to be detected by representational similarity analysis (RSA; Kriegeskorte et al., 2008).

2.2.3 Representation similarity analysis

Representational Similarity Analysis (RSA) is a computational technique developed by Kriegeskorte and colleagues (2008) that enables comparisons of different neural representational patterns across various measurement techniques, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and computational models. RSA measures the similarity or dissimilarity between representational patterns elicited in the brain under different conditions and transforms the results into a representational dissimilarity matrix (RDM), which provides a basis for comparing different experimental conditions or models based on the statistical distances between representations. The key advantage of RSA is that it facilitates the comparison of neural responses without directly aligning the raw activity across different individuals, measurement techniques, or experimental tasks. It becomes a very useful technique in detecting the brain activity patterns in response to different semantic categories and determining the extent to which the brain's representation of these categories is structured by their semantic properties. One particularly fruitful application of RSA has been in the investigation of animacy. Numerous studies have demonstrated that the brain represents animate and inanimate entities differently, and RSA has been a valuable tool in elucidating the neural underpinnings of this distinction with fMRI (Kriegeskorte et al., 2008) and with

magnetoencephalography (MEG)/EEG (Cichy et al., 2014; Cichy and Pantazis, 2017). Kriegeskorte et al. (2008) used RSA to compare neural representations of animate versus inanimate objects in the inferotemporal cortex, revealing that brain regions like the lateral occipital complex and posterior fusiform gyrus are particularly sensitive to animacy. Wehbe et al. (2014) also found that RSA could reveal patterns of neural similarity that corresponded to animacy distinctions in language comprehension, showing that brain areas involved in semantic processing (e.g., the anterior temporal lobes and angular gyrus) represent animate and inanimate entities differently. Anderson et al. (2016) found that RSA could compare neural responses during reading tasks with model-based predictions of animacy features, providing insight into how abstract semantic distinctions are encoded at both the linguistic and neural levels. Devereux et al. (2013) applied RSA to fMRI data collected while participants viewed words representing animate and inanimate objects. The results showed a clear dissociation between animate and inanimate word representations in the ventral temporal cortex, aligning with previous findings from object recognition studies but extending them to the domain of lexical semantics. In a more recent study, Wang et al. (2020) used RSA to identify neural patterns involved in predicting animacy features of upcoming nouns when the sentence context only restricts broad semantic features, such as animacy, rather than a specific word. In this study, verbs constrained the animacy of nouns (e.g., "caution" implied animate nouns, while "unfold" implied inanimate nouns). RSA revealed that prior to the appearance of nouns, neural activity patterns were more similar following verbs that implied animate nouns compared to those implying inanimate nouns, suggesting that the brain predicts coarse-grained semantic features beyond individual word prediction. However, one limitation of this approach is that neural similarity might reflect the meanings of the preceding context rather than the predicted words themselves, as context and predicted words are often correlated. In the current study, we tried to keep minimum context and directly compare EEG patterns after presenting classifiers without additional constraints. We specifically use RSA to examine the similarity between brain activity patterns following the classifiers,

right up until the nouns are presented. Since animate nouns tend to share more associative semantic features, while inanimate nouns have more distinct features (McRae et al., 2005; Zannino et al., 2006), neural activity patterns should exhibit greater similarity among animate nouns compared to inanimate nouns during noun presentation. The critical question is whether this similarity difference emerges before the nouns appear. If listeners are predicting the animacy of upcoming nouns, we expect greater similarity in neural patterns following animate-constraining classifiers compared to inanimate-constraining classifiers. We also analyze the similarity values post-noun presentation as a reference.

2.3 Article summary[1]

Prediction is proposed to be one of the fundamental mechanisms underlying language comprehension (Kuperberg and Jaeger, 2016). However, when the contexts are not constraining enough to predict specific words but can characterize a group of possible inputs, it remains unclear whether comprehenders can use such contextual constraints to pre-activate semantic features associated with the upcoming inputs. In this study, we ask whether the animacy constraints of classifiers can be utilized to predict semantic features associated with upcoming nouns. Chinese classifiers can provide animacy constraints on their head nouns, and we used EEG in combination with representational similarity analysis (RSA) to detect the pre-activation of animacy features. The basic assumption of RSA is that processing representationally similar items can cause similarities in brain activity patterns. Animate concepts share more intercorrelated semantic features than inanimate concepts (McRae et al., 1997; Zannino et al., 2006), and RSA has been used to distinguish between animate/inanimate concepts in fMRI (Kriegeskorte et al., 2008) and in MEG/EEG (Cichy & Pantazis, 2017; Wang et al., 2018). Therefore, we predict that if animacy features can be pre-activated regardless of the prediction of a specific word,

[1] The abstract of Article 1 has been submitted for the conference presentation in the 35th annual CUNY conference on human sentence processing.

then representational similarities should be greater following animate classifiers than inanimate classifiers before the onset of the nouns.

12 classifiers (6 animate-constraining and 6 inanimate-constraining) were each combined with 10 nouns to form 120 semantically matched classifier-noun pairs, and then they were recombined to form another 120 mismatched pairs (see appendix A for stimuli list). Participants (N=25) saw classifier-noun phrases in the word-by-word presentation mode (1000ms per word with 1000ms intervals) and were asked to judge whether the classifier-noun pairs were plausible by pressing a button. The experimental hypothesis rested on the assumption that animate nouns are more semantically similar to each other than inanimate nouns. Hence, we verified this assumption by quantifying and comparing semantic similarities of animate/inanimate nouns using a database HowNet. Additionally, to make sure that any difference in representational similarity reflects the pre-activation of animacy features of nouns rather than similarity associated with the animate/inanimate classifiers, we also verified that the two groups of classifiers match well in properties relevant to visual and linguistic processing including semantic similarity, visual complexity, and word frequency.

We calculated the similarity of EEG activity across 64 channels at each time point from the onset of the classifiers till the nouns within the animate/inanimate conditions. We then conducted paired t-tests consecutively with a step size of 10ms across the full time window to identify significant differences in similarity. No difference was found during the time windows of the two groups of classifiers. The animate classifier group showed greater representational similarity than the inanimate classifier group in the time window from -230ms till the onset of nouns ($p = 0.04$). The differences in similarities continue through the display of nouns for 300ms ($p = 0.01$).

RSA revealed that the similarity between neural activity following animate-constraining

classifiers was greater than that following inanimate-constraining classifiers, and critically, the effect of neural activity similarity emerged before the onset of nouns, indicating the pre-activation of the animacy feature of upcoming nouns. These findings provide neural evidence for the semantic prediction of features in language comprehension.

3 Article 2: Classifier animacy cues distinct gap expectations in head-final relative clause structure building in Mandarin Chinese

3.1 Research question: How detailed can prediction be?

Building upon Article 1, Article 2 also makes use of the animacy-constraining classifiers in Mandarin Chinese and to investigate whether the pre-activated animacy features can be further used to guide more detailed structural prediction in more sophisticated sentence constructions, i.e., filler-gap dependencies in relative clause constructions. By involving the prediction effects in relative clause constructions, this study is able to examine whether the details within the complex relative clause are predicted, thereby addressing the question of how detailed structural prediction can be during real-time language processing. Structural prediction refers to the anticipation of grammatical structures before encountering sufficient linguistic input to fully construct or confirm them. Understanding the granularity of these predictions sheds light on the general question of this thesis: what's the limit of prediction in real-time language processing? Moreover, detailed structural prediction requires immediate use of grammatical cues in the context, rather than broad grammatical knowledge based on frequency or probabilities, or heuristic parsing due to working memory constraints. Thus focusing on the detailed structural prediction can not only reveal how powerful predictive mechanisms can be in real-time language processing, but also help clarify how deeply grammatical knowledge is involved in real-time comprehension.

At a general level, theories of syntactic prediction differ in how fine-grained the predicted representations are. Within probabilistic parsing frameworks that use surprisal as a linking hypothesis (e.g., Hale, 2001; Levy, 2008), the degree of detail in prediction is determined by the grammar and parsing architecture over which the probability distribution is defined. Implementations based on relatively simple, unlexicalized PCFGs tend to generate coarser-grained expectations, for example, about phrase-closure and broad category transitions, whereas lexicalized and feature-rich grammars define surprisal over much

more articulated structures, including specific heads, subcategorization frames, and argument-role configurations. There are many empirical studies indicating that predictions can be remarkably detailed. For example, studies using self-paced reading and eye-tracking paradigms have shown that comprehenders anticipate specific verb argument structures (e.g., Altmann & Kamide, 1999). More impressive and detailed are the studies showing specific predictions based on world knowledge differences. For example, in sentences like “The man/girl will ride the…”, participants’ gaze patterns revealed that they anticipated specific objects (like a motorbike for a man or a carousel for a girl) before these objects were explicitly mentioned (Kamide & Altmann, 2003). Furthermore, ERP studies have found evidence for predictions at the level of morphosyntactic features, such as case markings (e.g., Wicha et al., 2004). These findings challenge the notion that predictions are limited to broad categories or probability-driven preferences and suggest that comprehenders may pre-construct highly specific syntactic structures depending on the specific cues that they pick up during the incremental processing of the previous context. Based on these different opinions, further research is needed to clarify the mechanisms underlying prediction and to determine the conditions under which detailed structural predictions are generated.

The general idea of this study is to investigate the comprehension’s ability to utilize subtle grammatical cues to construct detailed sentence structure in advance of the actual inputs. In Article 1 we captured the neurobiological evidence that animacy-constraining classifiers can elicit animacy feature pre-activation, now In Article 2, we focus on whether comprehenders can use the pre-activated animacy feature to further generate predictions to guide the gap search in relative clauses that are inserted in between the classifier and the head noun. The relative clause construction in Chinese is head-final, meaning the relative positions of the filler and the gap are reversed. Instead of processing the filler first and then searching for the gap, the gap proceeds the filler, making it more difficult to anticipate the existence of a relative clause and the gap site. Some studies focusing on

the filler-gap dependency processing in head-initial constructions have shown that animacy cues can modulate the detailed relative clause prediction, i.e., the gap site. However, existing evidence in filler-gap dependency processing in head-final constructions only shows how a mismatched classifier can help anticipate a relative clause without further exploring the factors affecting the gap site anticipation. Article 2 aims to take a step further to explore whether further detailed structure of the relative clause, i.e. the gap site, can be predicted. In the following Research Background, we will first review the well-established active gap search strategy in head-initial constructions like in English and how animacy can potentially modulate gap expectations. Then we will introduce what has been found about active prediction in head-final relative clause constructions like in Chinese.

3.2 Research background

3.2.1 Gap expectations: Active gap filling strategy and semantic modulators

One of the central challenges in real-time language comprehension is resolving syntactic dependencies, particularly those involving long-distance relationships, such as filler-gap dependencies. In many languages, sentences often contain a "filler" (a displaced constituent, like “the book” in (4)) that is associated with a "gap" (an unpronounced syntactic position, indicated as the underscore in (4)). Resolving the dependency relation is crucial for understanding the sentence, and comprehenders typically do so rapidly and efficiently. It comes naturally that predictive mechanisms are involved in constructing the dependency relation ahead of the actual presentation of the final part of such a long-distance dependency. The Active Gap Filling strategy (Fodor, 1978; Clifton and Frazier, 1989) is proposed to be the prominent mechanism to explain how comprehenders process such dependencies in real-time.

(4) The book [that I read _ yesterday] was fascinating.

Fodor (1978) first suggested that comprehenders actively search for gaps as soon as a filler is encountered. Early self-paced reading studies (e.g., Crain & Fodor, 1985) demonstrated slower reading times at regions where gaps were expected but absent, suggesting that comprehenders predict gaps and experience processing difficulty when their predictions fail. Self-paced reading studies, for example, Stowe (1986), provide compelling evidence for Active Gap Filling. They found that readers slow down when a gap site is unexpectedly filled with lexical material (e.g., "The nurse who the doctor argued that the patient met __"). These disruptions indicate an active search for the gap. An eye-tracking while reading study by Traxler and Pickering (1996) manipulated the plausibility between a displaced filler such as “the book” or “the city” and the verb as in “they talked about the book/city that the author wrote __…”, and found longer reading time at the verb when the filler is a semantically implausible object for the verb (city-wrote) compared with when the filler is a semantically plausible object (book-wrote). The observed plausibility mismatch effect demonstrated that comprehenders have built a sufficiently detailed structure, in which the filler takes on the thematic role of Theme of the verb and therefore confines its gap to the direct object position. That this effect could be observed as early as at the verb indicates that the active expectation for a gap site is rapidly initiated after the parser detects a filler. ERP studies further support this strategy. For example, ERP experiments have found P600 effects when a gap is absent or misaligned with expectations, indicating reanalysis and syntactic repair (Kaan et al., 2000).

While Active Gap Filling is fundamentally a syntactic strategy, it does not operate in isolation. Researchers have been working on how Active Gap Filling strategy can be modulated by other linguistic cues, for example, semantic plausibility, the degree to which a filler is semantically compatible with a potential gap site. Traxler and Pickering (1996) first focused on manipulating plausibility and tested how comprehenders posit gap sites. Their results showed that the comprehenders still posit gaps regardless of semantic plausibility but later showed disruption effects when the predicted gap sites turned out to

be semantically implausible. Other studies made use of animacy, a very fundamental and prominent semantic feature, to interact with active gap search and showed positive modulation effects. For example, the animacy of a filler has been shown to modulate the parser's expectation for relative clause type (Mak, Vonk & Schriefers, 2002, Traxler, Morris & Seely, 2005, Gennari & MacDonald, 2008). In a sentence completion task in Gennari and MacDonald's study (2008), they found that people tend to complete a relative clause as an object-gapped relative clause with given an inanimate filler compared with an animate filler, which they complete as a subject-gapped relative clause. In their self-paced reading experiment, they find that processing difficulty emerges as early as at the subject position in the relative clause when an animate filler noun phrase is provided. Lowder & Gordon (2014) also find a similar effect in eye-tracking. More recent work by Bovolenta and Husband (2023) shows evidence that subject-noun animacy can guide comprehenders to predict different verb phrase structures. For example, an inanimate subject predicts that the subject is derived, hence more object relative clauses will be predicted. Based on these results Gennari & MacDonald (2008, 2009) proposed the Production-Distribution-Comprehension account, arguing that an animate filler noun leads to an expectation of a subject gap while an inanimate filler noun induces a stronger expectation of an object gap. The central idea of expectation-based theories is that comprehenders dynamically adjust the likelihood of the upcoming linguistic inputs and generate up-to-date predictions based on the structure or features demonstrated in the previous inputs, highlighting the comprehenders' ability to utilize immediate grammatical knowledge to modulate sentence structural expectations. These results suggest that Active Gap Filling is not a purely syntax-driven process. Instead, it operates within a framework where syntactic predictions are dynamically adjusted based on semantic and contextual cues. This interaction prevents comprehenders from making implausible predictions, enhancing processing efficiency and reducing the need for costly reanalysis.

3.2.2 Chinese head-final relative clause

Evidence for the Active Gap Filling strategy has been observed in various Indo-European languages, including English, German, Dutch, etc. Phillips and Wagers (2007) argue that the Active Gap Filling strategy reflects a universal parsing sensitivity to linguistic dependencies. However, the prominence of Active Gap Filling may vary depending on the syntactic differences in the sentence constructions across languages, especially when the manifestations of filler-gap dependencies are different. In the head-initial relative clause constructions as we discussed in the previous section, the filler comes before the gap. However, there are also head-final relative clause constructions, e.g., in Chinese or Japanese, in which the relative order between filler and gap is reversed, see (5) as an example. This typological difference in filler-gap order may potentially affect how the parser perceives relative clause structures. In head-final relative clause constructions, active gap filling, which is assumed to be initiated by the filler in the previous studies, is not applicable as the filler is not available for the parser to initiate any gap search. Moreover, a relativizer that explicitly indicates the beginning of a relative clause is also absent in head-final relative clauses. Thus whether the parser can make any prediction about an upcoming relative clause becomes an intriguing question to investigate.

(5) A head-final relative clause construction in Mandarin Chinese:

[Jing cha zai yi yuan kan jian $\_\__i$]$_{RC}$ de na ge nv hai$_i$ zheng zai da dian hua

Police in hospital see REL that CL girl now make phone call

GAP FILLER

The girl who the police saw in the hospital walked into the classroom.

One sensible way to probe into this issue is to find what cues can indirectly indicate relative clause boundaries and whether they can be utilized by the parser to detect relative clauses in advance. The mismatch between linearly adjacent words can potentially indicate the boundary between the main clause and the embedded relative clause. For

instance, when an NP is modified by a classifier and a prenominal relative clause at the same time, the classifier can be put in front of the relative clause, as shown in (6a) and (6b). A classifier is semantically compatible with its head noun, meaning a classifier can only modify a specific type of nouns, and this close relation between a classifier and its head noun has been attested with much empirical evidence (Huettig et al., 2010, Zhang et al., 2012, Zhou et al., 2010). Thus, in cases like (6b) when a relative clause is inserted in between a classifier and a noun, the mismatch between the classifier ("ben" CL(book)) and the immediately following word ("peng you" "friend", a human noun) might potentially help comprehenders to detect an upcoming relative clause.

(6) Classifier immediately preceding its head noun:

a. [peng you tui jian] de zhe ben shu
Friend recommend REL this CL_{book} book
The book that a friend recommended.

Classifier dislocated from its head noun:

b. Zhe ben [peng you tui jian] de shu
This CL_{book} friend recommend REL book
The book that a friend recommended.

There is some evidence demonstrating that a semantically mismatched classifier can be effectively detected and used as a cue to predict relative clause structure in Mandarin Chinese (Hsu 2006, Wu et al, 2009, Chen et al, 2012; Wu et al, 2011, 2014, 2018). A self-paced reading study by Hsu in 2006 found the facilitation effect of classifier mismatch for predicting relative clause structure when presented with relative clause-biased discourse contexts, i.e., contexts which introduce two referents and create the need to use relative clause constructions to distinguish these two previously mentioned referents apart. Reasons were given by Hsu for why the facilitation effect of mismatched classifiers was

not effective in the absence of supporting contexts. The materials used in their studies were object-gapped relative clauses, thus the mismatched created was often between a classifier and a noun as shown in (7a) (e.g., …na wei jushi… "…that CL_{human} rock…"). It is a semantic plausibility mismatch rather than a syntactic categorical mismatch. Then one possible reason could be comprehenders' propensity to consider mismatched classifier-noun pairs in written texts as "typos". Furthermore, a corpus-based examination conducted by Wu and colleagues (2011) suggested that, for object-gapped relative clauses in Chinese, classifiers mostly appear in post-relative clause positions rather than dislocated to pre-relative clause positions. The situation is reversed for subject-gapped relative clauses; pre-relative clause classifiers occur much more frequently. This corpus finding might explain why Hsu's studies didn't find the facilitation effect of classifiers in object-gapped relative clauses without contexts. In subject-gapped relative clauses, the word following a pre-relative clause classifier is usually a verb or an adverb as shown in (7c) (e.g., …na kuai zazhong… "…that CL hit…"). This classifier-verb mismatch is more unambiguous and potentially more effective than a classifier-noun mismatch. Based on the results of the corpus study using revised materials, Wu and colleagues conducted a series of self-paced reading studies to examine whether classifier mismatch can serve as a reliable cue to facilitate relative clause prediction in isolated sentences. They found the facilitatory effect of mismatched classifiers in both subject-gapped relative clauses and object-gapped relative clauses without the support of contexts. And they also found that the reading time differences were numerically larger in subject-gapped relative clause conditions than those in object-gapped relative clause conditions, indicating the possibility that pre-relative clause classifiers might be more effective in subject-gapped relative clauses, i.e., classifier-verb mismatch than in object-gapped relative clauses, i.e., classifier-noun mismatch. Converging evidence also comes from the visual world eye-tracking paradigm by Wu and colleagues in 2014 showing that comprehenders can use mismatched classifier cues to anticipate a correct relative clause parse.

(7) Object-gapped relative clause with classifier:

a. na wei jushi zaidao de jizhe jingtide huangu sizhou.

That CL_{human} rock hit REL journalist cautiously look-about surroundings

The journalist that the rock hit looked about his surroundings cautiously.

Object-gapped relative clause without classifier:

b. jushi zadao de jizhe jingtide huangu sizhou.

rock hit REL journalist cautiously look-about surroundings

The journalist that the rock hit looked about his surroundings cautiously.

Subject-gapped relative clause with classifier:

c. na kuai zadao jizhe de jushi zhangzhe qingtai.

that CL_{rock} hit journalist REL rock grow moss

The rock that hit the journalist is covered with moss.

Subject-gapped relative clause without classifier:

d. na kuai zadao jizhe de jushi zhangzhe qingtai.

hit journalist REL rock grow moss

The rock that hit the journalist is covered with moss.

(Example sentence from Wu et al., 2009)

The studies mentioned above demonstrate how the parser is able to use mismatched classifiers to facilitate the processing of recognizing a relative clause domain. However, not much attention has been given to whether a classifier can also play a role in modulating the parser's expectancy of a gap after realizing the presence of a relative clause. How the filler-gap dependency is completed in head-final relative clauses remains

largely unknown. The active gap filling strategy found in head-initial relative clauses has not been tested much in head-final relative clauses due to the absence of filler in the pre-relative clause position. One underlying assumption of the active gap filling strategy when it was proposed is that it is a filler-driven process. The expectation of a relative clause and a gap at the earliest possible position is triggered after the parser perceives a filler and the motivation for such expectation is the working memory burden of maintaining a filler. Now that in head-final relative clause construction, a mismatched classifier can trigger an expectation of a relative clause, then the question is whether an expectation of a gap can also be triggered, or the processing of active gap search cannot be initiated in head-final relative clauses due to the absence of an actual filler. Dislocated pre-relative clause classifiers containing animate features allow us to test this issue. As we reviewed earlier, manipulating the animacy of filler NPs can modulate the parser's relative clause gap preference, so we wonder if animacy feature of a classifier can motivate an active prediction of gap sites.

3.3 Article summary[2]

Previous studies have suggested a predictive mechanism for relative clause (RC) processing in languages that have a head-final RC structure, like Japanese (Yoshida et al., 2004) and Mandarin Chinese (Hsu, 2006; Wu, 2009). However, it still remains unknown what type of information the parser utilizes to anticipate the structure of an upcoming RC and how detailed such structure building is before receiving information from the head noun directly. To address this, we investigated how the semantic information provided by different classifiers (CL) in Mandarin Chinese (human, non-human, general) guides the structure building of upcoming RCs.

[2] Abstracts of article 2 have been submitted for academic conferences including: the 26th Architectures and Mechanisms for Language Processing (AMLaP) and the 34th annual CUNY conference on human sentence processing.

Chinese “classifier + transitive verb” sequences are temporarily ambiguous between a subject-gapped RC and a (null subject) object-gapped RC construction. Although the parser is biased to adopt a subject RC analysis, semantic cues of a CL may be used to guide which of these two RC structures is initially adopted. Non-human (inanimate) CLs in particular may guide the parser away from a subject RC analysis since subjects are often Agents in canonical syntactic structure and inanimate CLs indicate that the head noun is unlikely to be an eligible subject for a subject RC. We predicted that this should facilitate the analysis of a null subject RC. With human and general CLs, the parser may be more likely to assume a subject gap and expect a noun to fill the object position. This predicts reading disruption upon encountering an unexpected relativizer and head noun. In a series of studies, CL type was manipulated to examine whether the parser uses CL type to predict the gap site in a head-final RC.

A sentence completion survey (N=439) was conducted online to investigate the parser’s bias for subject RC and null subject object RCs (see Appendix B: Experiment 1 for the stimuli list). The results suggest that the mismatch between a dislocated CL and the following verb guides the parser to an RC structure (88.7%) and the RC type is influenced by the CL type. Human CLs produce an overwhelming preference for subject-gapped RC (92.2%). General CLs also elicit a subject-gapped preference (71.4%). Non-human CLs, however, produce more object-gapped RC (85.9%).

Verbs and head nouns were selected based on the responses in the completion study and used as stimuli in an eye-tracking while reading experiment (N=42, see Appendix B: Experiment 2 for the stimuli list). Using general CL as baseline, results of linear mixed effect model show reading facilitation with non-human CL at the relativizer region in first fixation (Est=-12.24 ms, t=-2.399, $p<0.05$), first pass (Est=-14.17 ms, t=-2.545, $p<0.05$), go past (Est=-39.38 ms, t=-2.077, $p<0.05$) and total fixation (Est=-48.62 ms, t=-4.139, $p<0.001$). Human CL shows greater reading disruption compared with general CL in go-

past reading (Est=66.30 ms, t=3.499, p<0.01) and total fixation time (Est=46.59 ms, t=3.969, p<0.001). These effects are largely recapitulated at the head noun region. In the non-human CL condition, facilitation is significant in go-past reading (Est=-58.27 ms, t=-2.842, p<0.01) and total fixation (Est=-81.35 ms, t=-3.314, p<0.01). For human CL, disruption is significant in first-pass reading (Est=14.33 ms, t=2.326, p<0.05), go-past reading (Est=86.64 ms, t=4.310, p<0.001), and total fixation (Est=58.67, t=2.39, p<0.05).

We extended the results using self-paced reading, keeping the head nouns the same across different conditions by separately comparing non-human CL vs. general CL (N=43, see Appendix B: Experiment 3 for the stimuli list) and human CL vs. general CL (N=40). Both human and non-human conditions show reading disruptions at the verb (Est=35.08 ms, t=2.898, p<0.01; Est=30.37 ms, t=2.892, p<0.01), suggesting greater mismatch between the CLs and the verb. In human CL condition, disruptions continue in relativizer (Est=24.71 ms, t=2.413, p<0.05) and head noun (Est=37.16 ms, t=2.75, p<0.01) while in non-human CL condition, reading was facilitated at the relativizer (Est=-36.93 ms, t=-3.916, p<0.001) and the head noun (Est=-47.27 ms, t=-4.941, p<0.001).

The results indicate that the semantic properties of CLs can help the parser to make structural predictions in head-final RC processing before accessing the head noun. In particular, non-human CLs guide the parser away from the preferred subject-gapped RC structure, facilitating a null subject object-gapped analysis.

4 Article 3: Negative islands do not block active gap filling

4.1 Research question: Can pragmatic/presuppositional cues inhibit prediction?

Gap-site prediction in long-distance dependency is essentially a syntactic decision. It projects the clause types and clausal complexity, and the determination of a gap is mostly governed by the use of syntactic rules, for example, the number and type of syntactic heads (Dependency Locality Theory (DLT) by Gibson 1998, 2000), the distance between the head and the gap (also DLT by Gibson 1998, 2000), the argument structure of the verb (Pickering and Traxler, 2003; Frazier and Clifton, 1989; Boland et al., 1995) and syntactic constraints like island constraint (Stowe, 1986; Traxler and Pickering, 1996; Omaki and Schulz, 2011). However, factors from other linguistic levels can also interact with the syntactic rules in the process of generating gap expectations. Article 2 has demonstrated semantic cues can modulate gap expectations regardless of syntactic dispreference. Semantic cues like animacy features can direct the expectations of gap sites in gap-filler dependencies in Chinese head-final relative clauses. The demonstration of the interaction between semantic features and syntactic structure makes us eager to further probe other linguistic cues that might potentially affect syntactic structural expectations. Thus Article 3 targets an intriguing yet less understood element: the role of pragmatic cues. This study aims to address the question of whether pragmatic cues can modulate gap-site expectations in filler-gap dependency formation.

As noted earlier, island constraints have been known to have robust effects in prohibiting the formation of filler-gap dependencies across certain types of structures (e.g., Stowe, 1986; Traxler and Pickering, 1996). Violations of these constraints result in sentences that are perceived as ungrammatical or difficult to process. However, there are various types of island constraints and they can differ very much in terms of the source of the islandhood as well as the strength of the islandhood (detailed introduction in the following sections in Research Background). The so-called strong islands, where the extraction constraints are

caused by syntactic properties, have been the focus of psycholinguistic studies to demonstrate how grammatical rules affect real-time language processing. On the other hand, weak islands, where the extraction constraints are often selective and weaker, haven't been thoroughly accounted for. What's more, the real-time blocking effect of weak island constraints has, to my knowledge, not been tested. We target one particular weak island, negative islands, to investigate its interaction with active gap search in comparison with the blocking effect of a strong island constraint. The reason why we are interested in negative islands is that, this type of island constraint originated from pragmatic consideration rather than syntactic rules. To be more specific, some WH-words extractions that lead to degree or manner questions come with a contradictory presupposition when negation is present. Thus such a WH-word extraction is considered ungrammatical. This in turn might potentially block active gap filling when the fronted head is the constrained type. If negative islands have blocking effects for active gap filling comparable to the strong island constraints, it may suggest that presuppositional cues can be rapidly and effectively utilized by the comprehenders even in real-time language processing. If negative islands can not block active gap filling, it might indicate that there are different timings for linguistics cues from different levels to take effects on real-time processing decisions. It remains less unknown whether comprehenders can actively utilize pragmatic cues in real-time sentence processing, as pragmatic cues often involve broader world knowledge considerations. The effects of pragmatic cues are often tested in static or off-line grammaticality judgment tasks in which participants have enough time to engage in more throughout revisions. Those different types of island constraints, strong islands, and negative islands, provide us with a test ground to investigate the effects of pragmatic cues in comparison with syntactic cues on the real-time filler-gap dependency, and thus allow us to know more about how grammatical constraints from different linguistic levels works on structural predictions. A deeper understanding of this issue can shed light on core questions in psycholinguistics, such as how context interacts with linguistic structure and how cognitive resources are allocated during comprehension.

In the following research background, this study provides instruction on the island constraints and their blocking effects on active gap filling in filler-gap dependency processing. Then an overview of the negative island is presented to give a general idea of what we have known about negative island constraints. And finally, a research summary is given at the end.

4.2 Research background

4.2.1 Island constraints and the blocked filled-gap effect

Island constraints have long been studied within generative linguistics as limitations on possible syntactic movements. They are integral to understanding the relationship between sentence structure and the parsing mechanisms underlying sentence processing. The phenomenon of “island constraint” refers to restrictions on the possible syntactic dependencies that can be formed between the extracted element (normally, a wh-word) and its canonical position in the sentence, especially in the context of movement and gap-filler dependency relations. Although the formation of a dependency relation caused by Wh-movement is potentially unbound, as shown in (8a), the completion of such a dependency can still be restricted when certain types of constructions intervene. For example, when the canonical gap for the wh-word is located in a relative clause as in (8b), in a complex NP as in (8c), in a subject clause as in (8b), or in an adjunct clause as in (8e). These non-extraction domains are collectively referred to as “islands”, proposed by Ross (1967) in his seminal work on syntactic structures.

(8) a. Who did you hope that the candidate said that he admired __ ?

b. * Who did the candidate read a book that praised __?

c. * Who did the candidate read The Times' article about __?

d. * Who did the fact that the candidate supported __ upset voters in Florida?

e. * Who did the candidate raise two million dollars by talking to __?

(Examples from Phillips, 2006)

These islands as mentioned above provide syntactic environments where extraction is categorically forbidden, and they are referred to as strong islands. Violations of strong island constraints result in clear ungrammaticality. Strong islands have been widely recognized as universally robust constraints across languages, leading to hard violations of syntactic rules when extraction is attempted. In contrast, weak islands are environments that restrict extraction, but are not absolute constraints that block all extractions as strong islands do. Weak islands tend to impose selective constraints, depending on specific properties of the extracted element, normally with additional semantic requirements, such as why, how, or degree/quantity expressions. The reason for the selectivity is that weak island constraints often arise from contextual or semantic/pragmatic considerations that involve the validation with real world knowledge beyond literal inputs. For example, certain environments, particularly those that involve presuppositional or factive verbs (e.g., regret, realize), create weak island effects. For example, (9a) is generally judged as unacceptable because the extraction is not specific or definite. In the case of a more specific extraction like (9b), it becomes more acceptable. The extraction of "how" is more problematic compared to the extraction of "what", indicating that presuppositional islands are sensitive to the specificity or definiteness of the extracted element. Negative islands are another type of weak island which arises when negation in the sentence creates difficulties for certain types of extraction, especially for quantificational elements or adjuncts. For example, (9c) compares to (9d), the

extraction of degree questions (e.g., *how many people*) is blocked in the presence of negation, while the extraction of individual arguments (e.g., *who*) is allowed. This indicates that a weak island created by negation affects only specific types of dependencies, particularly those involving manners and degrees.

(9) a. * How do you regret that you didn't buy __?

b. What do you regret that you didn't buy __?

c. * How many people didn't John invite __?"

d. Who didn't John invite __?",

The fact that weak islands, no matter which type, only provide selective constraint, may suggest that mechanisms for the weak island constraint are different from strong islands. Strong islands are considered syntactic islands because they block any wh-element from moving out (who, what, how, why, etc.). This uniformity—where no type of extraction is allowed—suggests that the ban is a fundamental property of the syntactic structure itself. Early syntactic accounts for strong island formulated syntactic principles like Subjacency (Chomsky, 1973) or Barriers (Chomsky, 1986) which posited that strong islands arise from structural boundaries (e.g., NP, S) that cannot be crossed in a single wh-movement, capturing the robust ban on extraction. On the other hand, it has been a great challenge for any account that proposed that weak island constraint, such as factive island in (9a), or negative island in (9b) is caused by some syntactic property of negation or factive verbs (as in the Relativised Minimality of Rizzi,1990). Moreover, despite various semantic explanations proposed for weak islands, there is currently no coherent account that can comprehensively explain all types of weak island phenomena. Here in this study, we only focus on one specific type of weak island – the negative island caused by degree questions or manner questions, which is generally considered to be a pragmatic island constraint caused by presupposition violation (more detailed discussion see next section).

Island constraints are not only theoretical grammar rules. Many studies in psycholinguistics have found that island constraints can be imposed on the real-time sentence processor. Much research has been done to explore the island sensitivity of the real-time language parser. As the island constraint is closely related to the active gap filling strategy for filler-gap dependency formation, the disappearance of active gap filling effects is often taken as evidence for the effectiveness of an island constraint. A robust and well-known effect in active gap filling, the so-called “filled gap effect”, has long been used to snow how eager the parser is to fill the gap. As early as Frazier (1978) and Frazier & Clifton (1989), they provided influential insights into how readers and listeners anticipate a gap after encountering a wh-filler and exhibit processing slowdowns or misanalysis when that position turns out to be occupied by an overt constituent. Subsequent works by Stowe (1986) offered empirical evidence of this filled-gap effect in real-time parsing. Vice versa, the disappearance of the filled-gap effect has also been used by many researchers to demonstrate the island sensitivity of the real-time language parser, meaning that the parser stops searching for a gap inside of an island domain when it recognizes one. Also observed by Stowe (1986), the filled gap effect was not found in a subject-island domain as shown in (10a). The NP “Greg's”, as part of a complex subject NP, was read just as quickly as in its counterpart in (10b) where no filler-gap dependency is expected, suggesting that no gap site is posited after the preposition “about”.

(10) a. The teacher asked what the silly story about Greg’s older brother was supposed to mean.

b. The teacher asked what the team laughed about Greg’s older brother fumbling.

Ever since, the blocking effect of filled-gap has been used in many studies to demonstrate island sensitivity such as Pickering et al. (2015) in both their self-paced reading and eye-tracking experiment; Bourdages (1992) ‘s self-paced reading study in French relative

clause; McElree & Griffith (1998)'s speeded grammaticality judgment tasks in English relative clause and Yoshida et al. (2004) 's self-paced reading study in Japanese relative clause. Apart from the observation of the filled-gap effects using behavioral measurements, supporting evidence also comes from neurobiological measurements. Event-related potential (ERP) studies provide neurobiological evidence of how island constraints operate in real-time. ERP measures brain responses to stimuli, with certain components, like the P600, associated with syntactic processing difficulties. Kluender and Kutas (1993) found that sentences violating island constraints elicited stronger P600 effects, indicating that the brain treats these violations as syntactic errors, reinforcing the idea that island constraints play a crucial role in real-time language processing.

These findings suggest that the brain's real-time processing parser is sensitive to island constraints, to be more specific, strong island constraints, and actively prevents the formation of illicit dependencies, as evidenced by real-time disruptions in sentence processing. On the other hand, weak islands exhibit more subtle effects in real-time processing. In Kluender and Kutas (1993), they show unbounded dependencies elicit distinct ERP signatures (particularly a P600) when disrupted. Their results also revealed that these weak island violations trigger measurable neural responses indicative of syntactic difficulty. Although weak island effects were somewhat less pronounced than those observed for strong island violations (e.g., relative clauses, complex NPs), the ERP profile still showed that the parser was sensitive to the presence of the embedded wh-phrase. The amplitude and scalp distribution of the P600 suggested that, even in "weaker" contexts, the parser recognized these as illicit extraction sites. A more recent study by Villata et al. (2020), employing acceptability judgment tasks and maze-based self-paced reading, provides evidence that weak islands may influence active dependency formation in ways comparable to strong islands. Their investigation examined a range of different island types including both weak islands (e.g., whether-islands) and strong islands (e.g., adjunct islands and complex NP islands). Furthermore, eye-tracking experiments

conducted by Cokal and Sturt (2022) show that one type of weak island – the *whether* island exhibits a similar blocking effect for the filled-gap effect typically observed in strong islands, compared with non-island condition, suggesting that the parser has real-time sensitivity for both strong and weak islands. However, given the diversified nature of different types of weak islands, results for one type of weak island do not speak for the weak island as one unified category. Different types of weak islands might be originated from different levels of grammatical considerations. Whether-islands are still considered to be syntactic in nature with clear syntactic configurational boundaries.

The conflicts between the active gap search strategy and grammatical island constraints, and the subtlety of weak island effects allow us to dig deeper about questions like: how strong is the drive to actively search for a gap, and how strong is the grammatical constraint that prevents this drive from filling the gap immediately in long-distance relationships? Do the types of grammatical constraints matter when deciding whether to override the active gap search? Specifically, will a syntactic (strong island) constraint exert more influence than a semantic or pragmatic constraint (weak island) in preventing active gap filling? The overwhelming majority of psycholinguistic works on island constraints focus on the strong islands which provide unambiguous and frequently occurring syntactic constraints. Studies on the effect of weak islands are relatively rare as the weak island constraints are more subtle and obscure. This study targets at the effects of negative island constraints, which is one type of weak island constraint with a presuppositional nature.

4.2.2 Discussions in Negative islands

Negative islands, as mentioned earlier in the previous section, are a type of weak island referring to restrictions on certain kinds of movement in sentences containing negation. In particular, the negative island constraint limits the formation of wh-questions in contexts with negation, where extraction from a negated clause leads to ungrammaticality. For

example in (11a-d), extracting a which-question over negation (11b) is generally judged acceptable, but extracting a degree question over negation (11d) is unacceptable. Various accounts have been provided for negative islands in different levels of representation. A syntactic account has been proposed to explain negative islands by introducing negation as a potential antecedent governor that oversees wh-movement (Rizzi, 1990). Rizzi's syntactic account divides the wh-movement types into referential and non-referential expressions. A wh-expression assigned with a referential theta role can be extracted for movement while a wh-expression without a referential theta role, e.g., degree expressions, manner expressions, or measure expressions, cannot undergo extraction. Thus negation as a potential antecedent governor rejects a non-referential wh-phrase to be extracted over it but it cannot overrule a referential wh-extraction. Though Rizzi's account makes negative islands syntactically expressible, it does not necessarily suggest that the negative island constraint is completely subject to syntactic analysis. It is ultimately a semantic/pragmatic solution as noted by Kluender and Gieselman (2013) and also Rizzi in his later works (2003, 2004). Similar solutions (by examining the properties of wh-expressions) have also been given from a more semantic perspective. Szabolcsi and Zwarts (1993) suggest that the semantic properties of wh-expressions denote whether they can be legitimately extracted. Which- phrases, for example, demonstrate a characteristic of being an unordered element in a set of discrete individuals, while other wh-phrases, like manner expressions ("how"), degree expressions ("how tall"), etc., indicate a domain that is not individuated. Negation only permits extractions to happen with discrete individuals but rejects extractions with unindividuated phrases.

(11) a. Which man did John invite to the party?

b. Which man didn't John invite to the party?

c. How many children did the dog scare?

d. * How many children didn't the dog scare?

Apart from focusing on the properties of wh-expressions, an alternative semantic/pragmatic approach to explain negative island constraints pays attention to the principles for the formation of the question. One presupposition of question formation is that questions are subject to a principle of maximal informativity (Dayal, 1996), which states that all questions require one unique answer that is maximally informative, i.e., an answer set contains a true answer entailing all the other true ones (Fox & Hackl, 2006; Abrusán, 2011). (11b) meets this presupposition in the way that a most informative answer can be provided for this question, thus an extraction over negative does not cause a degraded acceptability for comprehenders. Degree questions like (11d) violate this presupposition. Because of the density of height, the true answer sets to this question are open intervals, so it is impossible to find a minimal or maximal height to answer this question that does not contain the actual height of John. As a result, it is judged to be unacceptable to form such a wh-extraction. There is also another claim proposing an existential presupposition borne by wh-questions supplement to the presupposition of maximal informativity (Comorovski, 1989; Kroch, 1989). If a uniquely identifiable referent that provides a pragmatically plausible answer as a contextually maximal informative answer is introduced in the larger discourse context, then a negative island constraint is rescued and such a question becomes acceptable. See examples in Kroch (1989) as in (12a) and (12b). If no specific amount of money that wasn't paid was given in the context, then subject to the maximal informativity principle, a question like (12a) is unacceptable. While if a unique amount of money is introduced into the previous context, or it can be inferred that such a specific amount is mentioned in the previous context, then the maximal informativity presupposition is considered fulfilled, and thus the question like (12b) becomes acceptable.

(12)

a. * How much didn't you pay?

b. How much didn't you pay that you were supposed to?

The presuppositional constraint accounts also explain why model verbs are able to rescue a negative degree question, which cannot be explained based on the accounts focusing on the properties of wh-expressions. Model verbs alleviate the presupposition requiring a unique answer, see the comparison between (13a) and (13b).

(13)

a. * How slow didn't he drive?

b. How slow shouldn't I be driving?

Shifting from grammatical constraint accounts for the unacceptability of negative islands, other studies go after a processing approach to understand the nature of negative island phenomena. The central claim of processing-based theories (e.g. Deane, 1991; Hofmeister & Sag, 2010; Kluender, 1991; Kluender, 1998; Kluender, 2004; Kluender & Kutas, 1993) is that the avoidance of wh-extractions out of certain domains might be a result of complex configurations that exceed the capacity of working memory system, rather than the violation of grammatical constraints. The processing costs are generated from a general processing perspective, that the unacceptability of negative questions is caused by individual factors that create accumulated processing difficulties, rather than by specific instantiation of grammatical knowledge. Gieselman and colleagues (2011) identified three factors that are known to increase processing difficulties in general: extraction, negation, and referentiality, and isolated each factor to see if that factor independently elicits differences in sentence acceptability with and without island contexts. In a series of acceptability tasks, they found robust effects of extraction types (subject-extractions vs. object-extractions) and of referent types (which-phrase vs. how many-

phrase) only in the presence of negation, while the effect of negation is prominent even in the absence of these other factors. They argue that negation is an extra-linguistic factor that modulates the acceptability of negative islands. However, it does not mean that island constraints can be reduced to processing factors that intervene in working memory load. There is also an increasing body of empirical evidence suggesting that the island effects remain significant even when general processing factors are controlled in the experimental designs (e.g., Sprouse, 2007; Sprouse, Wagers, & Phillips, 2012). What's more, their data come from offline acceptability tasks in which time-sensitive processing information is absent. The real-time status of negative island constraints has yet to be investigated. It remains unclear whether a negative Island constraint can be rapidly utilized by a real-time parser in sentence processing, or it is temporally ignored by the parser and is revised in a later stage. By addressing this issue, we can gain a better understanding of how the parser recognizes the properties that define an island domain, and further contribute to the discussion of the cues that can be effectively utilized by the predictive mechanism in language processing.

4.3 Article summary[3]

Constraints on long-distance dependencies arise not only from syntactic configurations but also from semantic/pragmatic considerations. Negative islands, a type of weak island, selectively constrain certain wh-dependencies that violate Dayal's (1996) maximal informativity presupposition on questions, i.e., that the answer set contains a true answer entailing all other true answers (Fox & Hackl, 2007; Abrusán, 2011). Negative degree questions like **How tall isn't John?* Are judged to be unacceptable because they ask for the minimal height interval that does not contain John's height, even though such an interval does not exist because the true answer set contains two mutually exclusive

[3] Abstracts of Article 3 have been submitted for academic conferences including 2nd Annual Conference on Experiments in Language Meanings; 29th Architectures and Mechanisms for Language Processing; 10th biennial meeting of Experimental Pragmatics.

subsets that do not entail one another, i.e. all intervals below John's height, (0, $height_{John}$), and all intervals above John's height, ($height_{John}$, ∞).

In general, strong island constraints are found to constrain long-distance dependency formation in real-time. Prior research demonstrates that comprehenders actively posit gaps for dependencies ahead of the input, revealed by slowdowns in reading times when those posited gaps are filled by other material, the so-called filled-gap effect. Stowe (1986) showed that comprehenders actively posit gaps for wh-phrases only in grammatical positions, demonstrating that filled-gap effects only emerge when gaps are grammatically licensed but not when they are grammatically inaccessible, e.g., inside subject islands. Further research has found that comprehenders respect wh-island constraints (Traxler & Pickering, 1996; Wagers & Phillips, 2009), reflecting the parser's rapid use of syntactic constraints to avoid positing illicit dependencies in real-time.

Whether comprehenders can use semantic/pragmatic constraints, such as negative islands, in real-time is unclear. Compared to syntactic constraints, it may take comprehenders more time to use presuppositions to block dependency formation, as calculating presupposition violations may be a more complex process. We examined whether negative islands are as effective as wh-islands at blocking illicit gaps in real-time. If comprehenders respect presuppositional constraints on dependencies, then we expect negative islands to be as effective as wh-islands in blocking a filled-gap effect. However, if comprehenders are unable to rapidly use presuppositional constraints to prevent comprehenders from positing illicit gaps, then we expect to see a filled-gap effect for negative islands, but not for wh-islands. Experiment 1 examined the offline acceptability of negative islands with (un)reduced relative clauses, to establish that comprehenders are sensitive to negative islands in offline complete reading. Experiment 2 (self-paced reading) and Experiment 3 (eye-tracking while reading) then investigated whether the filled-gap effect which is blocked in wh-islands is similarly blocked inside negative islands.

Experiment 1 is an acceptability judgment task (N=51, Items=24). We manipulated POLARITY (Positive, Negative) and STRUCTURE (No, Reduced, Unreduced RCs). See Appendix C: experiment 1 for the full stimuli list. While the presence of negation reduced acceptability overall (*Est.* = 0.37, *t* = 5.16, p < .001), there was a significant interaction of polarity with structure (*Est.* = 0.45, *t* = 4.23, p < .001). NoRC sentences (corresponding incrementally to a potential temporary gap in RRCs in Experiments 2 and 3) were rated much lower when negation was present (Est. = -1.64, t = -6.23, p < .001), suggesting that participants are sensitive to negative island constraints offline.

Experiment 2 is a self-paced reading experiment (N=63, Items=24). We manipulated the ISLAND type (No-, Neg-, Wh-Island), to examine whether comprehenders actively posit a gap inside islands (see Appendix C: Experiment 2&3, for the full stimuli list). A filled-gap effect emerged in the first spillover region between No-Island and Wh-Island conditions (*Est.* = 46.0, *t* = 3.12, *p* = .007), showing that Wh-Islands blocked dependency formation relative to No-Islands, but no significant difference was found between No-Islands and Neg-Islands (*Est.* = 28.8, *t* = 1.87, *p* = .157). Visually, however, Neg-islands appear to be intermediate between No-island and Wh-island conditions, suggesting some potential online sensitivity to negative island constraints. To examine this time course more closely, we followed up with an eye-tracking while reading study.

Experiment 3 is an eye-tracking while reading experiment (N=30; Item=24). We used the same experimental items from Experiment 2 to conduct an eye-tracking study, further probing into the time course of real-time negative island processing. At the critical adjective region, shorter reading times were found in wh-islands compared with no-islands in first-pass reading (Est. = 80.06, t = 2.60, p = .038), go-past reading (Est. = 145.9, t = 2.57, p = .029) and total reading (Est. = 242, t = 3.25, p = .009), reflecting the effectiveness of wh-islands in blocking illicit dependencies. However, no facilitation in reading time was

found in negative-islands compared with no-islands in either first-pass reading (Est. = -7.45, t = -0.25, p = .967), go-past reading (Est. = 13.6, t = 0.24, p = .969) or total reading (Est. = 115, t = 0.918, p = .633). These results suggest that negative islands are indeed unable to block active gap search. Furthermore, no visual trend was revealed, suggesting in self-paced reading trend was spurious.

In conclusion, although comprehenders are aware of negative islands offline, online results showed that they were unable to rapidly use this constraint to block active dependency formation. This asymmetry suggests that the effects of weak (semantic) islands take time to emerge, unlike strong (syntactic) islands which are more immediate.

5 General Discussion

5.1 Main findings

This project is centered around the topic of predictive mechanisms in real-time sentence processing, especially the ones involving long-distance dependencies. Many psycholinguistic studies have already established a solid foundation that prediction is involved in real-time language processing, however, there are still many unsolved questions regarding the capacity of language predictive mechanism: how actively comprehenders are engaged in prediction, how effectively comprehenders are able to utilize the linguistic cues for prediction and how detailed the prediction projection can be. Three experimental research projects were carried out to address these issues.

Article 1 focuses on the question of how actively comprehenders are engaged in making abstract predictions without highly constraining contact that elicits specific lexical prediction. To investigate whether animacy features can be pre-activated without relying on a highly constraining context, representational similarity analysis (RSA) was applied to EEG data. This study manipulated Chinese classifiers by their animacy constraints (animal-constraining v.s. inanimate-constraining) and paired them with nouns across three conditions: Congruent, Incongruent but Animacy-Matched, and Incongruent but Animacy-Mismatched. Behavioral results for the button pressing responses revealed that participants have faster responses (22 ms) for the Incongruent, Animacy-Mismatch condition than the Incongruent, Animacy-Matched condition, likely due to additional animacy violation. For the EEG results, a graded N400 effect was observed for both incongruent conditions, with larger negativity for the additional mismatch in animacy, which reflects the involvement of animacy in the processing of the semantic integration of nouns and preceding classifiers. Of greatest relevance to the study was the RSA result: the neural activity patterns following animate-constraining classifiers were more similar to each other compared to those following the inanimate-constraining classifiers. The

similarity effects were significant from ~200 ms before the onset of nouns. This result cannot be explained by lexical-semantic processing of the classifiers across conditions because they were matched well in semantic similarity structure and other linguistic properties. This finding is further confirmed by the matched ERP amplitudes across conditions during the presentation of classifiers. We suggest that the similarity difference reflects the pre-activation of semantic similarity of animal and inanimate nouns.

Article 2 focuses on the questions of how effectively comprehenders are able to utilize the linguistic cues for prediction and how detailed the prediction projection can be. Based on Article 1, Article 2 investigates whether comprehenders are able to further utilize the predicated animacy features to guide gap site anticipation in gap-filler dependency in Chinese head-final relative clauses based on semantic plausibility. An offline sentence completion task, an eye-tracking while reading experiment, and a self-paced reading experiment were conducted to investigate the role of animacy features in predicting gap sites in Chinese head-final relative clauses. The results of the sentence completion task show that a classifier-verb sequence elicits the production of different relative clause constructions depending on the animacy of the classifier. Inanimate classifiers lead to the generation of an object-gapped relative clause with the null subject which is otherwise dispreferred (Frazier, 1987; Kimball, 1973; Hsu, 2006). Animate and general classifiers both generated subject-gapped RCs, in line with the general preference. These findings suggest that classifier animacy cues are effective in guiding the construction of different relative clauses. Consistent with this completion study, the eye-tracking and self-paced reading experiments find real-time effects of classifier animacy on gap preferences. An inanimate classifier guides the parser to predict an object-gapped relative clause which facilitates the reading of a relativizer that occurs immediately after a transitive verb, while an animate classifier reinforces the prediction of a subject-gapped relative clause and creates reading disruption on a transitive verb-adjacent relativizer. The distinction between reinforcing a subject-gap relative clause analysis and prediction of an object-gap

relative clause analysis is made clearer in comparison to the general classifier which lacks animacy features and is consistent with either noun type. Both facilitation with inanimate classifiers and disruption with animate classifiers on transitive-verb-adjacent relativzers were found relative to the general classifier condition, suggesting that marked animacy features drove predictive structure-building.

Article 3 addresses the question of how effectively comprehenders are able to utilize linguistic cues for prediction. To start with, we confirmed that comprehenders are aware of negative island constraints offline in Experiment 1. The results show that, during off-line complete reading comprehension, negative degree questions are rated much lower than positive degree questions, suggesting that participants realized the violation of maximal informativity and thus considered negative degree questions to be unacceptable. An interaction between polarity and sentence structure suggested that relative clause island effectively blocked the scope of negation, rescuing the acceptability of negation with relative clauses. More importantly, using real-time experimental measures, Experiment 2 and Experiment 3 demonstrate the differences in time course when parsing different types of islands. We first confirmed the typical findings that strong islands can be utilized by the parser to block the active gap searching within island domains in both experiments. Then we demonstrate that the negative islands, on the other hand, cannot block the filled-fap effect caused by active gap filling and the reading patterns of negative islands are consistent with no-islands. Experiment 2 found the filled-gap effect in wh-island at the critical region, the first spillover region, and the second spillover region. While no effect was found in negative islands at the critical adjective region or any spillover region, though seemingly there was a trend towards faster recovery from the filled-gap effect at the spillover region. Using more fine-grained reading time measures, Experiment 3 replicates and solidifies the findings about the strong island's blocking effect in Experiment 2 and also further clarifies that the trend of numerically faster recovery from filled-gap effect at the spillover region in negative island condition was spurious. It is

evident by shorter reading times in strong island condition observed at the critical adjective region and spillover region with multiple reading time measures including first-pass reading time, go-past reading time and total reading time, and by no significant difference between negative islands and no-islands in any region in all reading measures.

In conclusion, Article 1 provides neurobiological evidence that Chinese classifiers can elicit animacy features prediction without specific lexical activation, meaning comprehenders are actively engaged in semantic prediction even in less-constraining context. Building on the finding in Article 1 and the robust observations of active gap search strategy during real-time sentence processing, Article 2 demonstrates that comprehenders can further utilize the classifier-elicited animacy features and the semantic plausibility of the animacy features for the thematic roles, to predict the plausible gap sites in the head-final relative clause structure in Chinese. Article 3 probes from a different perspective and targets at how effective are pragmatic cues in guiding active gap search in filler-gap dependencies in comparison with syntactic cues and the results suggest that presuppositional negative island constraint is not as effective as strong syntactic island constraint in blocking the active-gap search out of the constrained domains, meaning there might be different timings for linguistic cues from different levels to take effects, and pragmatic cues might take longer to compute and take effect on the real-time sentence processing.

5.2 Integrative Discussion

The three studies in this thesis touch upon different aspects about the scope and limitations of predictive mechanisms in real-time language processing involving long-distance dependency. Detailed discussion on the specific findings in each study can be found in the discussion section within each Article (see appendix). Here in this integrative discussion section, we highlight some points from a broader perspective.

Teasing apart feature-driven prediction and lexical-driven prediction

Articles 1 and 2 of this thesis collectively provide critical insights into the mechanisms of predictive processing by distinguishing feature-driven prediction from lexical-driven prediction. This distinction is fundamental in understanding what exactly triggers prediction in real-time language processing. As discussed earlier in the introduction, many experimental paradigms that elicit prediction in previous studies often heavily rely on the prediction of specific lexical items. When a highly constraining context elicits the prediction of a specific lexical item (e.g., “milk” as in “in the morning I would like to have a cup of coffee with sugar and __”), it also pre-activates relevant features of this lexical item. As a result, we cannot know if abstract features can be predicted and projected in the anticipatory processing independent from lexical activation. Another example is that it is always the head that triggers the active gap search in head-initial constructions (e.g., “book” as in “The book that the girl read __ is fascinating.”). The head is a lexical item in perfect form to provide all relevant linguistic cues for the comprehenders to actively search for the potential gap site. As a result, we cannot tease apart which feature(s) (e.g., number, gender, animacy, word class, affix) can function as the driving force in the search for a potential gap or whether it has to be a lexical item in full form to trigger the predictive search. Classifiers provide a suitable test ground to tease apart semantic features from lexical items and head-final relative clause construction which reverses the positional relative between the filler and the gap, allows for gap anticipation without the filler being present. The results from Article 1 and Article 2 demonstrate that an abstract feature alone, i.e., animacy feature, can be preactivated and can be utilized to trigger further structural prediction, i.e., gap site anticipation. It suggests that prediction can operate at a higher level of abstraction, activating categories (e.g., animate vs. inanimate) or broad semantic properties without linking to a specific word. It also adds evidence for language processing models that support feature-driven processing. Feature-driven models of language processing propose that linguistic comprehension and prediction are guided not just by specific lexical items but by abstract features (e.g., semantic categories, syntactic roles,

or morphosyntactic properties). These models highlight the flexibility and efficiency of using high-level information to constrain possible interpretations, especially in contexts where lexical specifics are ambiguous or unavailable. For example, Pickering & Garrod's (2013) Interactive Alignment Model argues that predictions are formed at multiple levels of representation (phonological, lexical, syntactic, and semantic). Feature-based predictions offer flexibility, allowing comprehenders to anticipate the general properties of upcoming input without committing prematurely to specific lexical items. This reduces the risk of costly reanalysis. Our results are also aligned with Surprisal Theory (Hale, 2001) which believes that predictions are based on probabilistic expectations. Abstract features like animacy reduce surprisal by constraining possible outcomes, even when specific lexical predictions are unavailable. Our results reinforce the idea that abstract features are a core component of these probabilistic computations.

Risk and modulation for predictive processing

Prediction in language comprehension trades efficiency against risk. By pre-constructing upcoming semantic features and syntactic structures, the parser can speed up interpretation, but at the cost of occasional mispredictions and reanalysis (Kuperberg & Jaeger, 2016). The results from Article 2 and Article 3 speak directly to how this balance is managed. On the one hand, animacy-constraining classifiers in Mandarin are exploited as reliable cues: comprehenders use them to anticipate semantic properties of upcoming nouns and to project gap sites in head-final relative clauses, thereby reducing processing effort in structurally complex configurations. On the other hand, comprehenders fail to consistently use presuppositional negative island constraints in English to modulate their predictions, even though these constraints are robust in off-line judgments. This dissociation suggests that the parser prioritizes cues that are both grammatically entrenched and locally available (such as classifier-encoded animacy), while weaker, context-dependent pragmatic cues are less likely to be recruited as real-time prediction triggers.

These findings bear on the relationship between grammar and the parser. Grammar, understood as the abstract system of rules and constraints, provides a rich set of potential guides for prediction, but the parser is a resource-limited, dynamic mechanism that implements only a subset of these constraints in real time. Article 2 shows a close alignment between grammar and parsing: classifier-noun agreement, animacy plausibility, and active gap search all shape both online measures and off-line judgments, indicating that certain grammatical constraints—especially those tied to locality and grammaticized features—are readily available to drive prediction. Article 3, by contrast, reveals a misalignment: negative islands show weak or delayed online effects despite clear off-line penalties. This pattern suggests that some presuppositional constraints are computed more slowly, or perhaps belong to a downstream interpretive component that evaluates complete propositions in context, rather than to the core feature system that licenses predictive structure building. In this sense, grammar is a necessary resource for prediction, but not all parts of grammar are equally accessible or equally weighted in real-time processing. The parser appears to rely most heavily on strong, early-available constraints—such as animacy and syntactic islands—while more global pragmatic constraints play a limited role in modulating prediction during incremental comprehension.

Theoretical Implications of Predictive Mechanisms

Collectively, the experimental findings from the three studies in this thesis allow us to take a more precise stance on the theoretical landscape of predictive processing mechanisms. As outlined and discussed in Chapter 2, we will evaluate the prediction triggers, prediction content and the algorithmic architecture of prediction. With respect to prediction cues, our results argue against the strict syntax-first view, while at the same time, also oppose the fully interactive view in which all cues are equally available to the parser. Article 1 and Article 2 show that animacy features encoded in classifiers, which carry strong semantic components, can be utilized as primary triggers for prediction. EEG representational similarity analysis results from Article 1 demonstrate that animacy information carried by

classifiers alone is sufficient to elicit different spatial patterns of brain activities related to the animacy features of the underspecified nouns before they appear. The results of head-final RC processing in Article 2 demonstrate that the same classifier animacy cue can actively guide predictions about the type of RC and the location of the gap inside an RC. These findings support interactive and expectation-based models in which syntactic and semantic cues are used jointly and very early, and they resonate in particular with architectures such as SPARSE and modified left-corner parsers, where semantic features are integrated into the grammatical feature bundles that license prediction. However, results from Article 3 suggest that not all cues are equally active as syntactic constraints and animacy constraints: strong syntactic island constraints robustly block active gap filling, while presuppositional negative island constraints fail to modulate the early active-gap search. This asymmetry suggests that there is still a hierarchy in the effective timing and weight of cues. A more appropriate view is that prediction is driven by interactions among an extended set of grammatical constraints, including core syntactic locality constraints and tightly grammaticized features such as classifier-encoded animacy, all of which are available early in processing to guide prediction. As for pragmatic constraints, even when they are indeed grammatically relevant, they are computed more slowly and do not reliably trigger predictive operations in time.

In terms of prediction content, the results provide clear evidence that prediction is neither restricted to coarse structural skeleton, nor always tied to highly specific lexical items. Article 1 shows that, in the absence of a deliberately highly constraining context, comprehenders can still predict abstract animacy features without committing to a particular noun. This finding suggests feature-driven prediction can be cleanly separated from lexical prediction. Then Article 2 demonstrates that these predicted features are not epiphenomenal: classifier-elicited animacy is used to shape the choice between subject- and object-gap RCs and to anticipate gap locations in a head-final RC construction, which is rather detailed structural configurations. It is precisely in line with the “feature-annotated

structural prediction" that SPARSE was designed to capture, positing underspecified predicted heads characterized by feature bundles and only later resolving them into fully specified categories and heads once licensing relations are satisfied. The present data thus lend empirical support to feature-based predictive architectures and extend expectation-based models by highlighting that what is predicted, in many cases, is a layered representation in which semantic feature bundles constrain a space of possible upcoming syntactic structures. Our findings also further inform our understanding of prediction commitment, that is, how "risky" the parser is in using its predictive capacities. Across the three studies, the human parser appears to be strikingly eager to pre-construct dependencies: animacy features are pre-activated in relatively weakly constraining classifier contexts, classifier plus verb and relativizer combinations are used to anticipate specific relative clause structures in Mandarin, and the parser continues to posit gaps in positions that violate negative island constraints in English. This profile aligns closely with the "hyper-active gap filler" view (Omaki et al., 2015) and with Crocker's characterization of the parser as "even more eager than a strictly incremental integrator", constructing anticipatory hypotheses beyond what is strictly required. Article 1 shows pre-activation of semantic features even when not strictly necessary for structure building. The classifier-driven animacy prediction is in addition to what a minimal integrative parser would need. Article 2 shows that the parser uses subtle classifier animacy cues to commit early to different RC types and gap positions in head-final structures, despite the well-known complexity and ambiguity of such configurations. Article 3, though suggesting that presuppositional constraints cannot modulate gap expectations, on the other hand, shows that the parser is eager to posit gaps inside negative islands, where such gaps will ultimately be judged as unlicensed. However, the eagerness of the parser does not necessarily mean hard commitments. The commitments are graded, and the parser can be constraint-sensitive. In Article 1, the predicted content remains at the level of coarse-grained animacy, rather than committing to a specific lexical form, which reduces the cost of prediction errors. In Article 2, the comparison between a human classifier and the

neutral general classifier, and the comparison between an inanimate classifier and the neutral general classifier, show intermediate patterns, meaning the structural commitment to the gap sites is graded. These findings portray the parser to be strongly predictive and risk-tolerant, but non-deterministic, consistent with probability-based models, expectation-based models, and with SPARSE's very soft commitment strategy in which multiple feature-compatible continuations remain available until licensing forces a more specific choice.

Taken together, our results point towards a refinement of the algorithmic architecture that underlies predictive mechanisms. Chapter 1 has already discussed that human sentence parsing must be at least partially top-down to maintain incrementality in head-final constructions. In the particular case of classifier-mismatch configuration as a cue for head-final RC construction in Mandarin Chinese, a left-corner parser lines up more cleanly to use the "classifier + verb mismatch" to predict an RC, though a sufficiently enriched top-down parser can be made to do it too. In the classifier-mismatch paradigm, the key moment is when the parser sees a classifier and then encounters a verb in a position where, under a simple NP continuation, it should be a noun. The CL + V sequence is incompatible with a simple CL + N NP construction, so this mismatch forces the parser to posit some extra structure, in which the verb is inside a relative clause modifying the unseen head. A top-down parser projects this structure from the top NP downward, based on phrasal rules and structural principles. When seeing a classifier, the parser already predicts that there will be a head noun showing up, and if the grammar or the structural principles implemented in the parser allow, it can optionally project an RC as well in the NP. But in this case, the classifier mismatch is not what drives the RC prediction. It might serve as a reanalysis cue when the parser only predicts a simple CL + N construction, but the CL + V mismatch is not the initial source of prediction. By contrast, a left-corner parser is built to use exactly this kind of local mismatch cue as the engine of prediction. It starts bottom-up from the left corner of a constituent and then projects upward. In the case of

the classifier mismatch in Mandarin Chinese, the parser first recognizes the classifier as a part of some NP projection. Then, when it encounters a verb, which is not compatible with the simple NP rule with CL + N, it moves up to a larger structure that can host a verb in that position, and such a hosting structure has to be an RC modifying the head N. So, in a left-corner architecture, the classifier mismatch triggers the parser to project an RC and anticipate a later head N. To further account for the observation that animacy features extracted from classifiers can act as early and powerful triggers for predicting RC type and gap location, a SPARSE-style feature-based left-corner parser fits our data the best. In SPARSE, when the parser encounters a classifier, it immediately creates a predicted head in the derivation, but this head is represented as a bundle of features rather than a fully specified category or lexical item. Our results show that animacy features should be encoded and plugged into the feature system that drives prediction and verification. When the parser encounters the verb, the mismatch suggests that the verb is the left corner of an RC, then the parser must choose how to expand the structure: toward a subject-gap RC or an object-gap RC. Because the predicted head already carries an animacy feature, the two RC schemata are evaluated relative to that feature bundle. A subject-gap RC is very compatible with a [+human] predicted head, so it is strongly favored after a human classifier; an object-gap RC is more compatible with a [- human] predicted head, so it becomes the preferred continuation after a non-human classifier. The gap location is therefore biased before the head noun appears, and the choice is explicitly driven by the classifier's feature contribution. At the same time, a feature-based SPARSE parser can also explain the asymmetry of island sensitivity between wh-islands and negative islands as observed in Article 3. A strong island like wh-island is a configuration where the grammar rules define that a [+wh] filler cannot license a gap across an embedded [+wh] clause boundary in the syntactic level. The feature bundle system of a SPARSE model works exactly on the level at which wh-dependency features of a wh-island are encoded. Thus, when the parser reaches a potential gap site inside a wh-island, the feature-checking mechanism will not allow the dependency to be formed, so no predicted gap is

attached there and thus no filled-gap effect. While the presuppositional constraint in a negative island cannot be reduced to local feature incompatibilities, like, “no [+wh] dependency across the [negation] boundary”, though negation should be the point when the comprehenders can realize the violation of the presuppositional requirement. The combination of wh-extraction and negation is better considered as a constraint on the interpretation of the whole clause. Thus in a SPARSE-style architecture, that means negative island constraints are not part of the feature bundles that drive prediction and licensing; they belong instead to a downstream interpretive/pragmatic component that evaluates complete propositions in context.

In light of these considerations, the thesis supports a theoretical picture in which prediction is best seen as implemented by a feature-guided, probabilistic left-corner architecture that operates over underspecified feature bundles in the style of SPARSE, and maintains graded, revisable commitments characteristic of expectation-based and constraint-based frameworks. Within this architecture, the three core aspects of prediction that motivated the thesis can now be more sharply characterized: prediction is licensed not only by syntax but by a restricted set of early-available localized features, especially animacy, while presuppositional constraints that work on the interpretive level of the whole clause are not included as prediction triggers. Prediction content is layered, ranging from feature-level properties to detailed structural configurations such as gap positions. And prediction commitment is eager in the sense of hyper-active dependence formation, but soft in that it remains probabilistic, feature-based, and sensitive to a hierarchy of constraints.

5.3 Limitations

There are some limitations of the studies, and they can mainly be categorized into the following three aspects. The first limitation is the sample size. Unlike recruiting participants for online tasks, the number of participants for a lab-based eye-tracking experiment was

limited by many external factors. The data collection for Article 1 and Article 2 happened during the Pandemic time. Therefore, there was a limited and restricted pool of participants available for the experiments. We made sure that the sample size for all experiments is comparable with previous studies which use similar experimental paradigms. However, we are looking for the utilization of subtle linguistic cues in sentences with complex structures, thus a larger sample size would be more helpful for the sake of getting robust effects.

The second general limitation is the potential influence of bilingualism. The data collection for Article 1 took place in the Institute of Psychology, Chinese Academy of Science, Beijing, and the data collection for the eye-tracking study in Article 2 took place in the Language and Brain Lab in the University of Oxford. Although the participants for Article 2 are all native speakers of Mandarin Chinese, given the fact that they all currently live in the UK, the influence of English is inevitable. It is not clear how bilingualism and the frequency of L1 and L2 usage influence on their relative clause processing and the utilization of classifier animacy cues.

Another potential concern is the narrow Region of Interest for the relativizer in Mandarin Chinese characters in Article 2. Unlike English, where words are alphabetic and separated by spaces, Mandarin Chinese uses logograms and most words in Mandarin Chinese are monosyllabic or disyllabic. It means word lengths in Mandarin Chinese are often short. Moreover, there are no clear spaces to indicate word boundaries, making it hard for word segmentations. The critical Interest Area in Article 2 was the relativizer region. The relativizer “de” in Mandarin Chinese is a monosyllabic functional particle. Due to its brevity, issues such as character skipping or inaccuracies in identifying fixation points might potentially affect eye-tracking data analysis. This issue is consistent for all experimental conditions, thus comparisons of reading time with each condition might be not greatly affected. However, the eye-tracking measures would be more accurate if there was a

proper way to solve this problem. Additionally, the visual complexity of Chinese characters, which can vary significantly between characters, may be one potential compound factor.

There are also some limitations concerning experimental design. In Article 1 and Article 2, we investigated the effect of animacy and selected some classifiers that modify distinctive animacy groups. However, animate/inanimate distinction is a spectrum rather than strictly categorical. Some noun groups that fall in the middle of the animacy hierarchy, such as plants (living but rarely taking Agent role in thematic position) or natural forces (like wind and storm, technically non-living but share some Agent features), are left unexamined. Article 3 also has a limited focus on specific types of negative islands. We only looked at negative islands that are caused by the maximal informativity presupposition on degree questions (e.g., how tall…) and number questions (e.g., how many…). The negation on these questions leads to infinite answers which violates the presupposition that there should be a maximal informative answer for the question. However, other negative island situations, even other types of weak islands are left unexplored. The results that we have found in Article 3 might not be generalizable for other types of negative islands. There is a further remark concerns the experimental design of Article 3. By manipulating the island types (WH-island, Neg-island, and No-island), we in fact manipulate two factors: negation and the presence of a WH-filler. Then a more effective and more commonly adopted experimental design should be a 2×2 factorial design with Polarity (Negative vs. Positive) and Island type (Present vs. Reduced), to better isolate the factors. Our initial design was indeed such a 2×2 design manipulating Polarity (Negative vs. Positive) and Island type (Present vs. Reduced). However, during piloting and initial data analysis, we found that the combination of a negation and a relative clause (e.g., How tall did Mary think the girl who was hoped not to be famous by her parents was before she went to college?) resulted in sentences that were syntactically and semantically more complex than the other conditions. This asymmetry made it difficult to construct natural and comparable items across all cells of the 2×2 design. To ensure

experimental balance and interpretability, we therefore opted for a 1×3 design (No-island vs. WH-island vs. Neg-island). This approach preserved the central contrast, which is to test whether sentential negation behaves as an island boundary, while avoiding potential confounds introduced by unnatural or structurally unbalanced materials. We believe this adjustment yields clearer results regarding the effect of negation on filler–gap dependency formation.

5.4 Future directions

One possible direction for future work might be to expand the semantic and pragmatic dimensions. In the present projects, the animacy feature is selected as one representation for semantic cues while future work could incorporate a wider range of semantic features (e.g., eventivity, instrumentality). Expanding the examination to a wider range of semantic features can give us a better idea about how these features interact and compete during comprehension and can further illuminate the hierarchical ordering of constraints and whether certain semantic or pragmatic cues are more readily integrated than others.

On the other hand, going deeper into the Animacy Hierarchy is another possibility. As mentioned in the limitations, animacy is not a binary category split between an animate group and an inanimate group. It functions as a spectrum, reflecting various degrees of animacy. Research on more categories in the middle of the Animacy Hierarchy can provide us with a more detailed picture of how animacy works. Moreover, animacy hierarchy influences the grammatical patterns in languages and it is even grammatically marked in some languages (e.g., Russian, Polish). However, different languages vary in how they draw the line between animate entities and inanimate entities. Investigations into whether people perceive animacy differently according to their native languages is also an interesting topic.

Another possible direction is to integrate the findings in these projects and use them to fine-tune the computational models that involve prediction. Computational models—

ranging from probabilistic Bayesian frameworks to large-scale neural network models—could be used to simulate and predict comprehension behaviors across experiments. By comparing model predictions with empirical data, researchers can identify which computational architectures best capture the interplay between semantic, syntactic, and pragmatic predictions. One potential challenge is that the current computational models of language processing might not be able to incorporate presuppositional grammatical constraints as investigated in Article 3. It's still debatable how to distinguish and recognize the extra-grammatical (including world knowledge) representations.

One more possibility is to explore how predictive strategies vary with individual differences, and language typological differences. Individual differences such as working memory capacity, language proficiency, or domain-general cognitive control abilities might reveal whether certain cognitive resources or experiences make comprehenders more adept at leveraging linguistic cues from specific levels in real-time comprehension. There are already some works focusing on individual differences in language comprehension. For example, the work by Ian Cunnings and Hiroki Fujita (2018) examines how individual cognitive capacities, such as working memory and language experience, influence sentence comprehension in both native and non-native speakers. Yadav and colleagues (2022) looked into individual differences in cue weighting during language processing of grammatical illusions. They found that participants varied dramatically in how they weighed different cues during sentence comprehension. Some individuals prioritized grammatical agreement cues while others relied heavily on word order cues. How these biased weighting of different linguistic cues affect predictive processing is a promising topic for future work.

Language typology can also play a role in cognitive resource allocation and parsing strategies. For example, for Chinese classifier cues in Article 1 and Article 2, work can be done to test parallel paradigms in languages with different morphosyntactic properties and

fewer (or more subtle) classifier systems (e.g., Japanese), to establish the universality of these findings. Such research would clarify to what extent the observed predictive strategies are grounded in universal cognitive mechanisms versus language-specific routines.

By pursuing these directions, we can deepen our understanding of the intricate, multi-level predictive machinery that underlies language comprehension. Over time, this will lead to a richer theoretical model that integrates various linguistic domains and cognitive processes into a cohesive framework of predictive language processing.

6 References

Abney, S. P. (1989). A computational model of human parsing. Journal of psycholinguistic Research, 18(1), 129-144.

Abrusán, M. (2011). Presuppositional and negative islands: A semantic account. Natural Language Semantics, 19(3), 257–321. https://doi.org/10.1007/s11050-011-9074-0

Altman, G., & Steedman, M. (1988). Interaction with context during human speech comprehension. Cognition, 30, 191-238.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. NeuroImage, 128, 44-53.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. Cognition, 21(1), 47-67.

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. Journal of memory and language, 34(6), 774-806.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. Language and Cognitive Processes, 26(3), 301-349.

Bourdages, J. S. (1992). Parsing complex NPs in French. In Island constraints: Theory, acquisition and processing (pp. 61-87). Dordrecht: Springer Netherlands.

Bovolenta, G., & Husband, E. M. (2023). Structural prediction during language comprehension revealed by electrophysiology: Evidence from Italian auxiliaries. Journal of Experimental Psychology: Learning, Memory, and Cognition, 49(1), 116.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. Journal of Cognitive Neuroscience, 10(1), 1-34. https://doi.org/10.1162/089892998563752

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. Empirical approaches to linguistic theory: Studies of meaning and structure, 43-62.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), A Festschrift for Morris Halle(pp. 232–286). Holt, Rinehart and Winston.

Chomsky, N. (1986). Barriers. MIT Press.

Cichy, R. M., & Pantazis, D. (2017). Multivariate pattern analysis of MEG and EEG: A comparison of representational structure similarity. NeuroImage, 158, 441-454.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. Nature Neuroscience, 17(3), 455-462. https://doi.org/10.1038/nn.3635

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences, 36(3), 181-204.

Clifton, C., & Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In G. Carlson & M. K. Tanenhaus (Eds.), Linguistic structure in language processing (pp. 273–317). Springer.

Cokal, D., & Sturt, P. (2022). The real-time status of strong and weak islands. PloS ONE, 17(2).

Comorovski, I. (1989). Discourse and the syntax of multiple constituent questions. Linguistic Inquiry, 20(4), 653–666.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), Natural language parsing: Psychological, computational, and theoretical perspectives (pp. 94–128). Cambridge University Press.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). Journal of experimental psychology: general, 132(2), 163.

Crocker, M. W. (1994). On the nature of the principle-based sentence processor. In Perspectives on Sentence Processing, edited by Jr. Clifton, Charles, Lyn Frazier, and Keith Rayner, 245–266. Lawrence Erlbaum Associates.

Crocker, M. (2002). Parsing as Prediction: Evidence from SVO and OVS constructions in German. In Proceedings of the 6th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002) (pp. 3-6).

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. Journal of psycholinguistic Research, 29(6), 647-669.

Cunnings, I., & Fujita, H. (2021). Quantifying individual differences in native and nonnative sentence processing. Applied Psycholinguistics, 42(3), 579-599.

Dayal, V. (1996). Locality in WH quantification: Questions and relative clauses in Hindi. Kluwer Academic Publishers.

Deane, P. D. (1991). Limits to attention: A cognitive theory of island phenomena. Cognitive Linguistics, 2(1), 1–63. https://doi.org/10.1515/cogl.1991.2.1.1

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

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

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. Journal of Neuroscience, 33(48), 18906-18916. https://doi.org/10.1523/JNEUROSCI.3809-13.2013

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. Journal of Cognitive Neuroscience, 10(1), 77-94. https://doi.org/10.1162/089892998563794

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. Cognition, 110(3), 293-321. https://doi.org/10.1016/j.cognition.2008.09.008

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. Psychological Science, 21(5), 629-634. https://doi.org/10.1177/0956797610367751

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2013). Are there mental lexicons? The role of semantics in lexical access. Psychological Review, 120(3), 453-479.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. Journal of Memory and Language, 41(4), 469-495. https://doi.org/10.1006/jmla.1999.2660

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. Linguistic Inquiry, 9(3), 427-473.

Fox, D., & Hackl, M. (2006). The universal density of measurement. Linguistics and Philosophy, 29(5), 537–586. https://doi.org/10.1007/s10988-006-9004-4

Frazier, L. (1987). Syntactic processing: evidence from Dutch. Natural Language & Linguistic Theory, 5(4), 519-559.

Frazier, L., & Clifton, C. (1986). The syntax of sentence structure. In Construal (pp. 26–53). The MIT Press.

Frazier, L., & Clifton, C. (1989). Successive cyclicity in the grammar and the parser. Language and Cognitive Processes, 4(2), 93-126. https://doi.org/10.1080/01690968908406359

Gelman, S. A., & Opfer, J. E. (2002). Development of the animate-inanimate distinction. In U. Goswami (Ed.), Blackwell handbook of childhood cognitive development (pp. 151-166). Blackwell.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. Journal of Memory and Language, 58(2), 161-187. https://doi.org/10.1016/j.jml.2007.07.004

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. Cognition, 111(1), 1-23.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1-76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), Image, language, brain (pp. 95-126). MIT Press.

Gieselman, S., Kluender, R., Caponigro, I., Fainleib, Y., LaCara, N., & Park, Y. (2013). Isolating processing factors in negative island contexts. In Proceedings of NELS (Vol. 41, pp. 233-246).

Giskes, A., & Kush, D. (2023). Abstract prediction of morphosyntactic features: Evidence from processing cataphors in Dutch. Glossa Psycholinguistics, 2(1).

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

Hale, J. (2003). The information conveyed by words in sentences. Journal of psycholinguistic research, 32(2), 101-123.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. Language, 86(2), 366-415. https://doi.org/10.1353/lan.0.0223

Hsu, C. C. N. (2006). Issues in head-final relative clauses in Chinese: Derivation, processing, and acquisition. University of Delaware.

Hsu, C. C., Tsai, S. H., Yang, C. L., & 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, Z., (2020). Classifier as a cue for structure building in head-final relative clause in Mandarin Chinese (M.Phil thesis). University of Oxford

Huang, Z., & Husband, E., M. (2021). Classifier as a cue for structure building in head-final relative clause in Mandarin Chinese. In 34th Annual CUNY Conference on Human Sentence Processing.

Huang, Z., & Husband, E., M. (2022). Negative islands do not block active gap filling. In 2nd Annual Conference on Experimental Language Meanings.

Huang, Z., Feng, C., & Qu, Q. (2023). Predicting coarse-grained semantic features in language comprehension: evidence from ERP representational similarity analysis and Chiense classifier. *Cerebral Cortex,* 33(13), 8312-8320

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. Acta psychologica, 137(2), 138-150.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458. https://doi.org/10.1038/nature17637

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. Journal of Memory and Language, 98, 1-11.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. Cognition, 50(1-3), 189-209.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. Cognitive science, 20(2), 137-194.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. Language and Cognitive Processes, 15(2), 159-201. https://doi.org/10.1080/016909600386084

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

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

Kazanina, N. (2017). Predicting complex syntactic structure in real time: Processing of negative sentences in Russian. Quarterly journal of experimental psychology, 70(11), 2200-2218.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, 105(31), 10687-10692. https://doi.org/10.1073/pnas.0802631105

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. Journal of cognitive neuroscience, 24(5), 1104-1112.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. Cognition, 2(1), 15-47.

Kimball, J. P. (1975). Predictive analysis and over-the-top parsing. In Syntax and Semantics volume 4 (pp. 155-179). Brill.

Kluender, R. (1991). Cognitive constraints on variables in syntax. Language, 67(2), 285–325. https://doi.org/10.2307/415093

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. In P. Culicover & L. McNally (Eds.), Syntax and Semantics (Vol. 29, pp. 241–279). Academic Press.

Kluender, R. (2004). Are subject islands subject to a processing account. In Proceedings of WCCFL (Vol. 23, pp. 475-499). Somerville, MA: Cascadilla Press.

Kluender, R., & Gieselman, S. (2013). What's negative about negative islands? A re-evaluation of extraction from weak island contexts. Experimental syntax and island effects, 186-207.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. Journal of Cognitive Neuroscience, 5(2), 196-214.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. Language and Cognitive Processes, 8(4), 573-633. https://doi.org/10.1080/01690969308407588

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2(4), 1–28. DOI: 10.3389/neuro.06.004.2008

Kroch, A. S. (1989). Amount quantification, referentiality, and long wh-movement. Manuscript. University of Pennsylvania.

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

Kwon, N., Sturt, P., & Liu, P. (2017). Predicting semantic features in Chinese: Evidence from ERPs. Cognition, 166, 433-446.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. Journal of Memory and Language, 61(3), 326-338. https://doi.org/10.1016/j.jml.2009.06.004

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. Nature reviews neuroscience, 9(12), 920-933.

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

Li X, Li X, Qu QQ. Predicting phonology in language compre- hension: evidence from the visual world eye-tracking task in mandarin Chinese. J Exp Psychol Hum Percept Perform. 2022:48(5): 531–547.

Li, C. N., & Thompson, S. A. (1989). Mandarin Chinese: A functional reference grammar. University of California Press.

Lowder, M. W., & Gordon, P. C. (2014). Effects of animacy and noun-phrase relatedness on the processing of complex sentences. Memory & cognition, 42, 794-805.

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological review, 101(4), 676.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain?. Trends in cognitive sciences, 15(3), 97-103.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. Journal of Memory and Language, 47(1), 50-68. https://doi.org/10.1006/jmla.2001.2837

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. Psychonomic Bulletin & Review, 23(4), 979-990. https://doi.org/10.3758/s13423-015-0842-3

McElree, B., & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(2), 432.

McRae K, De Sa VR, Seidenberg MS (1997) On the nature and scope of featural representations of word meaning. J Exp Psychol Gen 126:99–130.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior research methods, 37(4), 547-559.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language, 38(3), 283-312.

Ng, S., & Wicha, N. Y. (2014). Processing gap-filler dependencies in Chinese: What does it tell us about semantic processing?. Journal of memory and language, 74, 16-35.

Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. Studies in Second Language Acquisition, 33(4), 563-588. https://doi.org/10.1017/S0272263111000313

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

Phillips, C. (2006). The real-time status of island phenomena. Language, 795-823.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. In G. Gaskell (Ed.), Oxford handbook of psycholinguistics (pp. 739–756). Oxford University Press.

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

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and brain sciences, 36(4), 329-347.

Pickering, M. J., & Traxler, M. J. (2003). Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. Language and Cognitive Processes, 18(4), 469-503.

Pickering, M., Barton, S., & Shillcock, R. (2015). Unbounded dependencies, island constraints, and processing complexity. In Perspectives on sentence processing (pp. 199-224). Psychology Press.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. Journal of Memory and Language, 42(2), 168-182.

Pritchett, B. L. (1992). Grammatical competence and parsing performance. University of Chicago Press.

Rizzi, L. (1990). Relativized Minimality. MIT Press.

Rizzi, L. (2003). Relativized minimality effects. In M. Baltin & C. Collins (Eds.), The handbook of contemporary syntactic theory (pp. 89-110). Oxford, UK: Blackwell.

Rizzi, L. (2004). Locality and left periphery. In A. Belletti (Ed.), Structures and beyond: The cartography of syntactic structures (Vol. 3, pp. 223-251). Oxford, UK: Oxford University Press.

Roark, B. (2001). Robust probabilistic predictive syntactic processing. arXiv preprint cs/0105019.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. Behavioral and Brain Sciences, 31(6), 689-714.

Rosenkrantz, D. J., & Lewis, P. M. (1970, October). Deterministic left corner parsing. In 11th Annual Symposium on Switching and Automata Theory (swat 1970) (pp. 139-152). IEEE.

Ross, J. R. (1967). Constraints on variables in syntax (Doctoral dissertation, Massachusetts Institute of Technology). MIT Press.

Rullmann, H. (1995). Maximality in the Semantics of Wh-constructions. University of Massachusetts Amherst.

Schneider, D. A. (1999). Parsing and incrementality. University of Delaware.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319.

Sprouse, J. (2007). A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. Doctoral dissertation, University of Maryland.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. Language, 88(1), 82–123. https://doi.org/10.1353/lan.2012.0004

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(2), 425.

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. Language and Cognitive Processes, 1(3), 227-245. https://doi.org/10.1080/01690968608407062

Szabolcsi, A., & Zwarts, F. (1993). Weak islands and an algebraic semantics for scope taking. Natural language semantics, 1(3), 235-284.

Szabolcsi, A., & Zwarts, F. (1993). Weak islands and an algebraic semantics for scope taking. Natural Language Semantics, 1(3), 235–284. https://doi.org/10.1007/BF00263545

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. Journal of Memory and Language, 68(4), 297-314.

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. Language and cognitive processes, 4(3-4), SI211-SI234.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632–1634.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. Journal of Memory and Language, 35(4), 454-475. https://doi.org/10.1006/jmla.1996.0025

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. Journal of Memory and Language, 53(2), 204-224.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of memory and language, 33(3), 285-318.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. Trends in Cognitive Sciences, 5(6), 244-252. https://doi.org/10.1016/S1364-6613(00)01651-X

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & 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 Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. International Journal of Psychophysiology, 83(2), 176-190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Villata, S., Tabor, W., & Sprouse, J. (2020, March). Gap-filling in syntactic islands: Evidence for island penetrability from the maze tasks. In The 33st Annual CUNY Conference on Human Sentence Processing, Amherst (US) (pp. 19-21).

Vissers, C. T. W., Chwilla, D. J., & Kolk, H. H. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. Brain Research, 1106(1), 150-163.

Wagers, M. W., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. Journal of Linguistics, 45(2), 395-433. https://doi.org/10.1017/S0022226709005726

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. elife, 7, e39061.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: evidence from MEG and EEG representational similarity analysis. Journal of Neuroscience, 40(16), 3278-3291.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 233-243).

Wicha, N. Y. Y., Moreno, E. M., & 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. https://doi.org/10.1162/0898929041920487

Wu, F., Kaiser, E., & Andersen, E. (2009). The effect of classifiers in predicting Chinese relative clauses. In the Proceedings of the Western Conference on Linguistics (WECOL). Davis: University of California. Proceedings online: http://wecol. ucdavis. edu.

Wu, F., Kaiser, E., & Vasishth, S. (2018). Effects of early cues on the processing of Chinese relative clauses: Evidence for experience-based theories. Cognitive science, 42, 1101-1133.

Wu, F., Luo, Y., & Zhou, X. (2014). Building Chinese relative clause structures with lexical and syntactic cues: Evidence from visual world eye-tracking and reading times. Language, Cognition and Neuroscience, 29(10), 1205-1226.

Wu, Fuyun. Frequency issues of classifier configurations for processing Mandarin object-extracted relative clauses: A corpus study. (2011): 203-207.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation. Open Mind, 6, 1-24.

Yoshida, M. (2004). Relative clause prediction in Japanese. In 17th Annu. CUNY Sentence Process. Conf., University of Maryland, College Park, MD, 2004.

Zannino GD, Perri R, Pasqualetti P, Caltagirone C, Carlesimo GA (2006) Analysis of the semantic representations of living and nonliving concepts: a normative study. Cogn Neuropsychol 23:515–540.

Zannino, G. D., Perri, R., Pasqualetti, P., Caltagirone, C., & Carlesimo, G. A. (2006). (Category-specific) semantic deficit in Alzheimer's patients: The role of semantic distance. Neuropsychologia, 44(1), 52-61.

Zhang, Y., Zhang, J., & 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.

Zhou, X., Jiang, X., Ye, Z., Zhang, Y., Lou, K., & Zhan, W. (2010). Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study. Neuropsychologia, 48(6), 1551-1562.

7 Appendices

APPENDIX A: EXPERIMENTAL ITEMS FOR ARTICLE 1

Classifier type	Classifier subtype	Classifier	Noun	Noun_type
animate	human	一位	专家	match
animate	human	一位	法官	match
animate	human	一位	少女	match
animate	human	一位	战友	match
animate	human	一位	老伯	match
animate	human	一位	前辈	match
animate	human	一位	嘉宾	match
animate	human	一位	军医	match
animate	human	一位	孕妇	match
animate	human	一位	富商	match
animate	human	一名	学生	match
animate	human	一名	患者	match
animate	human	一名	员工	match
animate	human	一名	司机	match
animate	human	一名	游客	match
animate	human	一名	乘客	match
animate	human	一名	考生	match
animate	human	一名	强盗	match
animate	human	一名	厨师	match
animate	human	一名	佣人	match
animate	human	一代	英雄	match
animate	human	一代	大师	match
animate	human	一代	领袖	match
animate	human	一代	元帅	match
animate	human	一代	帝王	match
animate	human	一代	伟人	match
animate	human	一代	军师	match
animate	human	一代	圣贤	match
animate	human	一代	传人	match
animate	human	一代	宗师	match
animate	animal	一只	花猫	match
animate	animal	一只	老鼠	match
animate	animal	一只	老虎	match
animate	animal	一只	天鹅	match

animate	animal	一只	蝴蝶	match
animate	animal	一只	熊猫	match
animate	animal	一只	狐狸	match
animate	animal	一只	苍蝇	match
animate	animal	一只	孔雀	match
animate	animal	一只	小鸟	match
animate	animal	一窝	小猫	match
animate	animal	一窝	老鼠	match
animate	animal	一窝	蚂蚁	match
animate	animal	一窝	燕子	match
animate	animal	一窝	小鸡	match
animate	animal	一窝	兔子	match
animate	animal	一窝	麻雀	match
animate	animal	一窝	耗子	match
animate	animal	一窝	毒蛇	match
animate	animal	一窝	喜鹊	match
animate	animal	一头	狮子	match
animate	animal	一头	骆驼	match
animate	animal	一头	小猪	match
animate	animal	一头	母猪	match
animate	animal	一头	大象	match
animate	animal	一头	奶牛	match
animate	animal	一头	水牛	match
animate	animal	一头	毛驴	match
animate	animal	一头	母牛	match
animate	animal	一头	犀牛	match
inanimate	natural	一朵	棉花	match
inanimate	natural	一朵	玫瑰	match
inanimate	natural	一朵	鲜花	match
inanimate	natural	一朵	牡丹	match
inanimate	natural	一朵	白云	match
inanimate	natural	一朵	梅花	match
inanimate	natural	一朵	桃花	match
inanimate	natural	一朵	乌云	match
inanimate	natural	一朵	蘑菇	match
inanimate	natural	一朵	浪花	match
inanimate	natural	一声	呼唤	match
inanimate	natural	一声	叹息	match

inanimate	natural	一声	咳嗽	match
inanimate	natural	一声	叫喊	match
inanimate	natural	一声	呐喊	match
inanimate	natural	一声	枪响	match
inanimate	natural	一声	闷雷	match
inanimate	natural	一声	呼喊	match
inanimate	natural	一声	尖叫	match
inanimate	natural	一声	巨响	match
inanimate	natural	一滩	污水	match
inanimate	natural	一滩	鲜血	match
inanimate	natural	一滩	积水	match
inanimate	natural	一滩	墨水	match
inanimate	natural	一滩	血迹	match
inanimate	natural	一滩	烂泥	match
inanimate	natural	一滩	泥浆	match
inanimate	natural	一滩	泥水	match
inanimate	natural	一滩	淤泥	match
inanimate	natural	一滩	雪水	match
inanimate	artifact	一份	工作	match
inanimate	artifact	一份	报纸	match
inanimate	artifact	一份	公告	match
inanimate	artifact	一份	提案	match
inanimate	artifact	一份	快餐	match
inanimate	artifact	一份	心意	match
inanimate	artifact	一份	表格	match
inanimate	artifact	一份	问卷	match
inanimate	artifact	一份	简历	match
inanimate	artifact	一份	试卷	match
inanimate	artifact	一件	事情	match
inanimate	artifact	一件	礼物	match
inanimate	artifact	一件	行李	match
inanimate	artifact	一件	大衣	match
inanimate	artifact	一件	衬衫	match
inanimate	artifact	一件	案子	match
inanimate	artifact	一件	礼服	match
inanimate	artifact	一件	外套	match
inanimate	artifact	一件	毛衣	match
inanimate	artifact	一件	雨衣	match

inanimate	artifact	一块	蛋糕	match
inanimate	artifact	一块	石头	match
inanimate	artifact	一块	砖头	match
inanimate	artifact	一块	豆腐	match
inanimate	artifact	一块	饼干	match
inanimate	artifact	一块	黑板	match
inanimate	artifact	一块	玻璃	match
inanimate	artifact	一块	石碑	match
inanimate	artifact	一块	肥皂	match
inanimate	artifact	一块	宝石	match
animate	human	一位	苍蝇	filler
animate	human	一位	孔雀	filler
animate	human	一位	小鸟	filler
animate	human	一位	耗子	filler
animate	human	一位	毒蛇	filler
animate	human	一位	蘑菇	filler
animate	human	一位	血迹	filler
animate	human	一位	淤泥	filler
animate	human	一位	试卷	filler
animate	human	一位	事情	filler
animate	human	一名	麻雀	filler
animate	human	一名	毛驴	filler
animate	human	一名	母牛	filler
animate	human	一名	天鹅	filler
animate	human	一名	蝴蝶	filler
animate	human	一名	雪水	filler
animate	human	一名	白云	filler
animate	human	一名	桃花	filler
animate	human	一名	棉花	filler
animate	human	一名	玫瑰	filler
animate	human	一代	喜鹊	filler
animate	human	一代	狮子	filler
animate	human	一代	水牛	filler
animate	human	一代	小猪	filler
animate	human	一代	母猪	filler
animate	human	一代	鲜花	filler
animate	human	一代	快餐	filler
animate	human	一代	心意	filler

animate	human	一代	石头	filler
animate	human	一代	砖头	filler
animate	animal	一只	伟人	filler
animate	animal	一只	军师	filler
animate	animal	一只	厨师	filler
animate	animal	一只	司机	filler
animate	animal	一只	嘉宾	filler
animate	animal	一只	叫喊	filler
animate	animal	一只	工作	filler
animate	animal	一只	积水	filler
animate	animal	一只	泥浆	filler
animate	animal	一只	泥水	filler
animate	animal	一窝	少女	filler
animate	animal	一窝	英雄	filler
animate	animal	一窝	领袖	filler
animate	animal	一窝	军医	filler
animate	animal	一窝	孕妇	filler
animate	animal	一窝	烂泥	filler
animate	animal	一窝	报纸	filler
animate	animal	一窝	行李	filler
animate	animal	一窝	雨衣	filler
animate	animal	一窝	蛋糕	filler
animate	animal	一头	乘客	filler
animate	animal	一头	考生	filler
animate	animal	一头	强盗	filler
animate	animal	一头	战友	filler
animate	animal	一头	老伯	filler
animate	animal	一头	礼物	filler
animate	animal	一头	公告	filler
animate	animal	一头	案子	filler
animate	animal	一头	礼服	filler
animate	animal	一头	宝石	filler
inanimate	natural	一朵	元帅	filler
inanimate	natural	一朵	帝王	filler
inanimate	natural	一朵	佣人	filler
inanimate	natural	一朵	骆驼	filler
inanimate	natural	一朵	老鼠	filler
inanimate	natural	一朵	墨水	filler

inanimate	natural	一朵	外套	filler
inanimate	natural	一朵	毛衣	filler
inanimate	natural	一朵	问卷	filler
inanimate	natural	一朵	饼干	filler
inanimate	natural	一声	员工	filler
inanimate	natural	一声	小鸡	filler
inanimate	natural	一声	游客	filler
inanimate	natural	一声	花猫	filler
inanimate	natural	一声	犀牛	filler
inanimate	natural	一声	提案	filler
inanimate	natural	一声	黑板	filler
inanimate	natural	一声	玻璃	filler
inanimate	natural	一声	表格	filler
inanimate	natural	一声	肥皂	filler
inanimate	natural	一滩	圣贤	filler
inanimate	natural	一滩	传人	filler
inanimate	natural	一滩	老虎	filler
inanimate	natural	一滩	老鼠	filler
inanimate	natural	一滩	宗师	filler
inanimate	natural	一滩	石碑	filler
inanimate	natural	一滩	尖叫	filler
inanimate	natural	一滩	简历	filler
inanimate	natural	一滩	衬衫	filler
inanimate	natural	一滩	乌云	filler
inanimate	artifact	一份	专家	filler
inanimate	artifact	一份	法官	filler
inanimate	artifact	一份	蚂蚁	filler
inanimate	artifact	一份	燕子	filler
inanimate	artifact	一份	大师	filler
inanimate	artifact	一份	呐喊	filler
inanimate	artifact	一份	咳嗽	filler
inanimate	artifact	一份	浪花	filler
inanimate	artifact	一份	闷雷	filler
inanimate	artifact	一份	呼喊	filler
inanimate	artifact	一件	前辈	filler
inanimate	artifact	一件	兔子	filler
inanimate	artifact	一件	小猫	filler
inanimate	artifact	一件	大象	filler

inanimate	artifact	一件	奶牛	filler
inanimate	artifact	一件	豆腐	filler
inanimate	artifact	一件	巨响	filler
inanimate	artifact	一件	污水	filler
inanimate	artifact	一件	鲜血	filler
inanimate	artifact	一件	梅花	filler
inanimate	artifact	一块	富商	filler
inanimate	artifact	一块	学生	filler
inanimate	artifact	一块	患者	filler
inanimate	artifact	一块	熊猫	filler
inanimate	artifact	一块	狐狸	filler
inanimate	artifact	一块	大衣	filler
inanimate	artifact	一块	枪响	filler
inanimate	artifact	一块	牡丹	filler
inanimate	artifact	一块	呼唤	filler
inanimate	artifact	一块	叹息	filler

APPENDIX B: EXPERIMENTAL ITEMS FOR ARTICLE 2

Experiment 1: sentence completion

喜欢 like
General CL 那个喜欢____
Human CL 那位喜欢____
Non-human CL 那首喜欢____

讨厌 hate
General CL 那个讨厌____
Human CL 那位讨厌____
Non-human CL 那门讨厌____

珍藏 cherish
General CL 那个珍藏____
Human CL 那位珍藏____
Non-human CL 那段珍藏____

碰倒 knock down
General CL 那个碰倒____
Human CL 那位碰倒____
Non-human CL 那盆碰到____

要看 to see
General CL 那个要看____
Human CL 那位要看____
Non-human CL 那部要看____

担心 worry
General CL 那个担心____
Human CL 那位担心____
Non-human CL 那件担心____

推荐 recommend
General CL 那个推荐____
Human CL 那位推荐____
Non-human CL 那本推荐____

偏爱 favor
General CL 那个偏爱____
Human CL 那位偏爱____
Non-human CL 那首偏爱____

在意 mind
General CL 那个在意____
Human CL 那位在意____
Non-human CL 那件在意____

带来 bring
General CL 那个带来____
Human CL 那名带来____
Non-human CL 那捆带来____

带走 take
General CL 那个带走____
Human CL 那位带走____
Non-human CL 那瓶带走____

学会 learn
General CL 那个学会____
Human CL 那名学会____
Non-human CL 那项学会____

要问 to ask
General CL 那个要问____
Human CL 那位要问____
Non-human CL 那件要问____

惦记 think about
General CL 那个惦记____
Human CL 那位惦记____
Non-human CL 那本惦记____

闻到 smell
General CL 那个闻到____
Human CL 那位闻到____
Non-human CL 那股闻到____

听到 hear
General CL 那个听到____
Human CL 那位听到____
Non-human CL 那条听到____

转发 repost
General CL 那个转发____
Human CL 那名转发____
Non-human CL 那条转发____

捡到 pick up
General CL 那个捡到____
Human CL 那位捡到____
Non-human CL 那张捡到____

抹去 erase
General CL 那个抹去____
Human CL 那名抹去____
Non-human CL 那段抹去____

打破 break
General CL 那个打破____
Human CL 那位打破____
Non-human CL 那片打破____

推倒 push down
General CL 那个推倒____
Human CL 那名推倒____
Non-human CL 那堵推倒____

清空 empty
General CL 那个清空____
Human CL 那名清空____
Non-human CL 那张清空____

吃了 eat PERF
General CL 那个吃了____
Human CL 那位吃了____
Non-human CL 那块吃了____

卖了 sale PERF
General CL 那个卖了____
Human CL 那位卖了____
Non-human CL 那支卖了____

切开 cut
General CL 那个切开____
Human CL 那名切开____
Non-human CL 那块切开____

倒掉 pour
General CL 那个倒掉____
Human CL 那名倒掉____
Non-human CL 那桶倒掉____

扔掉 throw
General CL 那个扔掉____
Human CL 那名扔掉____
Non-human CL 那张扔掉____

浪费 waste
General CL 那个浪费____
Human CL 那名浪费____
Non-human CL 那碗浪费____

举起 lift up
General CL 那个举起____
Human CL 那名举起____
Non-human CL 那根举起____

找到 find
General CL 那个找到____
Human CL 那位找到____
Non-human CL 那本找到____

Experiment 2: Eye tracking

Sentences	Condition
那个捡到的宝贝现在被过路人占为己有了。	general
那名捡到的男孩现在被好心人送到了医院检查身体。	human
那张捡到的银行卡现在被好心人放到了当地派出所。	non-human
那个扔掉的布娃娃已经变得脏兮兮的了。	general
那名扔掉的孤儿已经变得脏兮兮的了。	human
那张扔掉的草稿纸已经变得皱巴巴的了。	non-human
那个讨厌的小狗一直叫个不停。	general
那位讨厌的领导要召开一个紧急会议。	human
那门讨厌的课程终于就要上完了。	non-human
那个卖掉的小猫实际上是品种稀罕的波斯猫。	general
那名卖掉的小孩实际上有先天性心脏病。	human
那款卖掉的手表实际上是十八世纪的古董。	non-human
那个喜欢的玩偶仍然摆在他的床头。	general
那位喜欢的姑娘仍然每天去集市上卖菜。	human
那本喜欢的书仍然放在他的枕头边上。	non-human
那个惦记的味道明天就可以回家吃到了。	general
那位惦记的姑娘明天就可以见到了。	human
那场惦记的比赛明天就要正式开始了。	non-human
那个切开的西瓜是刚刚从冰箱里拿出来的。	general
那名切开的受害者是刚刚在郊区树林里发现的。	human
那块切开的蛋糕是刚刚从烤箱里拿出来的。	non-human
那个抹去的痕迹给警察带来了调查困难。	general
那位抹去的伟人仍然活在每一个人的心中。	human
那段抹去的记忆就像一阵烟一样地飘散了。	non-human
那个捕获的动物被猎人扛上了卡车。	general
那名捕获的犯人被警察关进了看守所。	human
那条捕获的大鱼被渔民冻进了冷冻柜。	non-human
那个举起的杠铃足足有五公斤重。	general
那名举起的路人狠狠地又摔倒了地面上。	human
那根举起的棍子狠狠地敲在了小李的后背上。	non-human
那个推倒的自行车压倒了一大堆瓦片。	general
那名推倒的病人躺在地上不断的哀嚎。	human
那堵推倒的墙象征着勇敢的人们对自由的向往。	non-human
那个推荐的商品受到了广大消费者的喜爱。	general
那位推荐的学者获得了老师和学生的一致推崇。	human
那部推荐的电影获得了奥斯卡最佳电影奖。	non-human

Experiment 3a: self-paced reading: non-human vs general classifier

Sentences	Condition
那块偷走的怀表其实是十八世纪的古董。	non-human
那个偷走的怀表其实是十八世纪的古董。	general
那盒带走的药片里面藏着走私的毒品。	non-human
那个带走的药片里面藏着走私的毒品。	general
那场惦记的比赛明天就要正式开始了。	non-human
那个惦记的比赛明天就要正式开始了。	general
那瓶丢弃的饮料晚上被保洁人员收走了。	non-human
那个丢弃的饮料晚上被保洁人员收走了。	general
那张发现的照片是十年前我们一起拍的。	non-human
那个发现的照片是十年前我们一起拍的。	general
那条忽略的线索才是破案的关键。	non-human
那个忽略的线索才是破案的关键。	general
那枚捡到的硬币被小孩用水冲洗干净了。	non-human
那个捡到的硬币被小孩用水冲洗干净了。	general
那根举起的棍子狠狠地敲在了小李的后背上。	non-human
那个举起的棍子狠狠地敲在了小李的后背上。	general
那款卖掉的手机是今年的限量款。	non-human
那个卖掉的手机是今年的限量款。	general
那段抹去的记忆就像一阵烟一样地飘散了。	non-human
那个抹去的记忆就像一阵烟一样地飘散了。	general
那盆碰倒的水仙已经快要枯萎了。	non-human
那个碰倒的水仙已经快要枯萎了。	general
那首偏爱的歌表达了他对家乡的思念。	non-human
那个偏爱的歌表达了他对家乡的思念。	general
那块切开的蛋糕是刚刚从烤箱里拿出来的。	non-human
那个切开的蛋糕是刚刚从烤箱里拿出来的。	general
那张扔掉的信纸已经变得皱巴巴的了。	non-human
那个扔掉的信纸已经变得皱巴巴的了。	general
那门讨厌的课程终于就要上完了。	non-human
那个讨厌的课程终于就要上完了。	general
那条捕获的大鱼很快就被渔民冻进了冷冻柜。	non-human
那个捕获的大鱼很快就被渔民冻进了冷冻柜。	general
那堵推倒的墙象征着勇敢的人们对自由的向往。	non-human
那个推倒的墙象征着勇敢的人们对自由的向往。	general
那部推荐的电影获得了奥斯卡最佳电影奖。	non-human
那个推荐的电影获得了奥斯卡最佳电影奖。	general
那本喜欢的书仍然放在他的枕头边上。	non-human
那个喜欢的书仍然放在他的枕头边上。	general
那篇选中的文章被刊登在了当地报纸上。	non-human
那个选中的文章被刊登在了当地报纸上。	general

那首遗忘的歌谣歌颂的是一段凄美的爱情。	non-human
那个遗忘的歌谣歌颂的是一段凄美的爱情。	general
那件在意的事一直都萦绕在他的脑海里。	non-human
那个在意的事一直都萦绕在他的脑海里。	general
那本找到的笔记带给了他创作的灵感。	non-human
那个找到的笔记带给了他创作的灵感。	general
那段珍藏的回忆一直保存在我们的心底。	non-human
那个珍藏的回忆一直保存在我们的心底。	general
那根剪断的绳子正好符合门框的长度。	non-human
那个剪断的绳子正好符合门框的长度。	general
那片打破的玻璃上面有血迹。	non-human
那个打破的玻璃上面有血迹。	general
那项分配的任务十分棘手，需要小明连夜完成。	non-human
那个分配的任务十分棘手，需要小明连夜完成。	general
那棵砍断的大树已经活了将近三百年了。	non-human
那棵砍断的大树已经活了将近三百年了。	general
那碗浪费的面汤只能倒进垃圾桶里。	non-human
那个浪费的面汤只能倒进垃圾桶里。	general
那门取消的考试不会影响到学生的绩点。	non-human
那个取消的考试不会影响到学生的绩点。	general

Experiment 3b: self-paced reading: human vs general classifier

Sentences	Condition
那位找到的证人可能不愿意作证。	human
那个找到的证人可能不愿意作证。	general
那名捕获的犯人被法官判决死刑。	human
那个捕获的犯人被法官判决死刑。	general
那名丢弃的孩子已经浑身脏兮兮的了。	human
那个丢弃的孩子已经浑身脏兮兮的了。	general
那名发现的受害者全身都是伤口。	human
那个发现的受害者全身都是伤口。	general
那位忽略的路人才是案情的关键。	human
那个忽略的路人才是案情的关键。	general
那名激怒的工人十分凶狠地瞪了老板一眼。	human
那个激怒的工人十分凶狠地瞪了老板一眼。	general
那名捡到的男孩生怕自己再一次被送走。	human
那个捡到的男孩生怕自己再一次被送走。	general
那个开除的记者现在正在准备投诉报社。	general
那位开除的记者现在正在准备投诉报社。	human
那名卖掉的小孩已经找到了亲生父母。	human
那个卖掉的小孩已经找到了亲生父母。	general
那位抹去的伟人仍然被工人阶级铭记。	human
那个抹去的伟人仍然被工人阶级铭记。	general
那位撞倒的老人已经昏迷不醒。	human
那个撞倒的老人已经昏迷不醒。	general
那个偏爱的学生数学考试又得了第一。	human
那名偏爱的学生数学考试又得了第一。	general
那名扔掉的孤儿患有先天心脏疾病。	human
那个扔掉的孤儿患有先天心脏疾病。	general
那位讨厌的领导明天要召开一个紧急会议。	human
那个讨厌的领导明天要召开一个紧急会议。	general
那位提拔的员工更加卖力地工作了。	human
那个提拔的员工更加卖力地工作了。	general
那名偷走的小孩被卖到了山区。	human
那个偷走的小孩被卖到了山区。	general
那名推倒的学生躺在地上不断哀嚎。	human
那个推倒的学生躺在地上不断哀嚎。	general
那位推荐的学者最近出了一本著作。	human
那个推荐的学者最近出了一本著作。	general
那位喜欢的女孩每天课间都在读书。	human
那个喜欢的女孩每天课间都在读书。	general
那名想念的姑娘马上就要离开这个小镇了。	human
那个想念的姑娘马上就要离开这个小镇了。	general

那位选中的用户收到了神秘礼物。	human
那个选中的用户收到了神秘礼物。	general
那个邀请的嘉宾感到十分荣幸。	human
那名邀请的嘉宾感到十分荣幸。	general
那位遗忘的老人依然平静地过着自己的生活。	human
那个遗忘的老人依然平静地过着自己的生活。	general
那位在意的警官没有放弃调查这个案子。	human
那个在意的警官没有放弃调查这个案子。	general
那名解雇的员工开始四处投简历。	human
那个解雇的员工开始四处投简历。	general
那位戏弄的宾客并没有感到愤怒。	human
那个戏弄的宾客并没有感到愤怒。	general
那名出卖的同事感到十分愤怒。	human
那个出卖的同事感到十分愤怒。	general
那位提拔的员工下个月就要涨工资了。	human
那个提拔的员工下个月就要涨工资了。	general
那名激怒的少年握紧了拳头。	human
那个激怒的少年握紧了拳头。	general
那名逗乐的客人高兴地鼓起了掌。	human
那个逗乐的客人高兴地鼓起了掌。	general

APPENDIX C: EXPERIMENTAL ITEMS FOR ARTICLE 3

Experiment 1: offline grammaticality judgment

Condition1	Condition2	Sentence
gap	negative	How tall did Mary think the girl hoped not to be?
gap	positive	How tall did Mary think the girl hoped to be?
reduced	negative	How tall did Mary think the girl hoped not to be famous was?
reduced	positive	How tall did Mary think the girl hoped to be famous was?
relative	negative	How tall did Mary think the girl who was hoped not to be famous was?
relative	positive	How tall did Mary think the girl who was hoped to be famous was?
gap	negative	How successful did Lily think the attorney hoped not to be?
gap	positive	How successful did Lily think the attorney hoped to be?
reduced	negative	How successful did Lily think the attorney hoped not to be wealthy was?
reduced	positive	How successful did Lily think the attorney hoped to be wealthy was?
relative	negative	How successful did Lily think the attorney who was hoped not to be wealthy was?
relative	positive	How successful did Lily think the attorney who was hoped to be wealthy was?
gap	negative	How famous did Fred think the singer hoped not to be?
gap	positive	How famous did Fred think the singer hoped to be?
reduced	negative	How famous did Fred think the singer hoped not to be thinner was?
reduced	positive	How famous did Fred think the singer hoped to be thinner was?
relative	negative	How famous did Fred think the singer who was hoped not to be thinner was?
relative	positive	How famous did Fred think the singer who was hoped to be thinner was?
gap	negative	How objective did John think the editor hoped not to be?
gap	positive	How objective did John think the editor hoped to be?
reduced	negative	How objective did John think the editor hoped not to be strict was?
reduced	positive	How objective did John think the editor hoped to be strict was?
relative	negative	How objective did John think the editor who was hoped not to be strict was?
relative	positive	How objective did John think the editor who was hoped to be strict was?
gap	negative	How frank did Andy think the interviewer hoped not to be?
gap	positive	How frank did Andy think the interviewer hoped to be?
reduced	negative	How frank did Andy think the interviewer hoped not to be constructive was?
reduced	positive	How frank did Andy think the interviewer hoped to be constructive was?
relative	negative	How frank did Andy think the interviewer who was hoped not to be constructive was?
relative	positive	How frank did Andy think the interviewer who was hoped to be constructive was?
gap	negative	How popular did Mary think the kid wished not to be?
gap	positive	How popular did Mary think the kid wished to be ?
reduced	negative	How popular did Mary think the kid wished not to be quiet was?
reduced	positive	How popular did Mary think the kid wished to be quiet was?
relative	negative	How popular did Mary think the kid who was wished not to be quiet was?
relative	positive	How popular did Mary think the kid who was wished to be quiet was?

gap	negative	How muscular did James think the student wished not to be?
gap	positive	How muscular did James think the student wished to be ?
reduced	negative	How muscular did James think the student wished not to be talkative was?
reduced	positive	How muscular did James think the student wished to be talkative was?
relative	negative	How muscular did James think the student who was wished not to be talkative was?
relative	positive	How muscular did James think the student who was wished to be talkative was?
gap	negative	How happy did Chris think the woman wished not to be?
gap	positive	How happy did Chris think the woman wished to be?
reduced	negative	How happy did Chris think the woman wished not to be married was?
reduced	positive	How happy did Chris think the woman wished to be married was?
relative	negative	How happy did Chris think the woman who was wished not to be married was?
relative	positive	How happy did Chris think the woman who was wished to be married was?
gap	negative	How brave did Lily think the girl wished not to be?
gap	positive	How brave did Lily think the girl wished to be?
reduced	negative	How brave did Lily think the girl wished not to be popular was?
reduced	positive	How brave did Lily think the girl wished to be popular was?
relative	negative	How brave did Lily think the girl who was wished not to be popular was?
relative	positive	How brave did Lily think the girl who was wished to be popular was?
gap	negative	How adaptable did Robin think the journalist wished not to be?
gap	positive	How adaptable did Robin think the journalist wished to be ?
reduced	negative	How adaptable did Robin think the journalist wished not to be incisive was?
reduced	positive	How adaptable did Robin think the journalist wished to be incisive was?
relative	negative	How adaptable did Robin think the journalist who was wished not to be incisive was?
relative	positive	How adaptable did Robin think the journalist who was wished to be incisive was?
gap	negative	How slim did John think the doctor wished not to be?
gap	positive	How slim did John think the doctor wished to be?
reduced	negative	How slim did John think the doctor wished not to be gentle was?
reduced	positive	How slim did John think the doctor wished to be gentle was?
relative	negative	How slim did John think the doctor who was wished not to be gentle was?
relative	positive	How slim did John think the doctor who was wished to be gentle was?
gap	negative	How innocent did Jack think the defendant claimed not to be?
gap	positive	How innocent did Jack think the defendant claimed to be?
reduced	negative	How innocent did Jack think the defendant claimed not to be guilty was?
reduced	positive	How innocent did Jack think the defendant claimed to be guilty was?
relative	negative	How innocent did Jack think the defendant who was claimed not to be guilty was?
relative	positive	How innocent did Jack think the defendant who was claimed to be guilty was?
gap	negative	How responsible did Brian think the president claimed not to be?
gap	positive	How responsible did Brian think the president claimed to be?
reduced	negative	How responsible did Brian think the president claimed not to be healthy was?
reduced	positive	How responsible did Brian think the president claimed to be healthy was?

relative	negative	How responsible did Brian think the president who was claimed not to be healthy was?
relative	positive	How responsible did Brian think the president who was claimed to be healthy was?
gap	negative	How heavy did Lily think the player claimed not to be?
gap	positive	How heavy did Lily think the player claimed to be?
reduced	negative	How heavy did Lily think the player claimed not to be strong was ?
reduced	positive	How heavy did Lily think the player claimed to be strong was ?
relative	negative	How heavy did Lily think the player who was claimed not to be strong was ?
relative	positive	How heavy did Lily think the player who was claimed to be strong was ?
gap	negative	How poor did Jason think the child claimed not to be?
gap	positive	How poor did Jason think the child claimed to be?
reduced	negative	How poor did Jason think the child claimed not to be intelligent was?
reduced	positive	How poor did Jason think the child claimed to be intelligent was?
relative	negative	How poor did Jason think the child who was claimed not to be intelligent was?
relative	positive	How poor did Jason think the child who was claimed to be intelligent was?
gap	negative	How convincing did Judy think the officer expected not to be?
gap	positive	How convincing did Judy think the officer expected to be?
reduced	negative	How convincing did Judy think the officer expected not to be present was?
reduced	positive	How convincing did Judy think the officer expected to be present was?
relative	negative	How convincing did Judy think the officer who was expected not to be present was?
relative	positive	How convincing did Judy think the officer who was expected to be present was?
gap	negative	How smart did James think the child expected not to be?
gap	positive	How smart did James think the child expected to be?
reduced	negative	How smart did James think the child expected not to be polite was?
reduced	positive	How smart did James think the child expected to be polite was?
relative	negative	How smart did James think the child who was expected not to be polite was?
relative	positive	How smart did James think the child who was expected to be polite was?
gap	negative	How positive did Amy think the interviewee expected not to be?
gap	positive	How positive did Amy think the interviewee expected to be?
reduced	negative	How positive did Amy think the interviewee expected not to be hard-working was?
reduced	positive	How positive did Amy think the interviewee expected to be hard-working was?
relative	negative	How positive did Amy think the interviewee who was expected not to be hard-working was?
relative	positive	How positive did Amy think the interviewee who was expected to be hard-working was?
gap	negative	How dependable did Adam think the manager expected not to be?
gap	positive	How dependable did Adam think the manager expected to be?
reduced	negative	How dependable did Adam think the manager expected not to be bossy was?
reduced	positive	How dependable did Adam think the manager expected to be bossy was?
relative	negative	How dependable did Adam think the manager who was expected not to be bossy was?
relative	positive	How dependable did Adam think the manager who was expected to be bossy was?
gap	negative	How pessimistic did Kate think the lawyer expected not to be?
gap	positive	How pessimistic did Kate think the lawyer expected to be?

reduced	negative	How pessimistic did Kate think the lawyer expected not to be patient was?
reduced	positive	How pessimistic did Kate think the lawyer expected to be patient was?
relative	negative	How pessimistic did Kate think the lawyer who was expected not to be patient was?
relative	positive	How pessimistic did Kate think the lawyer who was expected to be patient was?
gap	negative	How nervous did Kate think the candidate claimed not to be?
gap	positive	How nervous did Kate think the candidate claimed to be?
reduced	negative	How nervous did Kate think the candidate claimed not to be successful was?
reduced	positive	How nervous did Kate think the candidate claimed to be successful was?
relative	negative	How nervous did Kate think the candidate who was claimed not to be successful was?
relative	positive	How nervous did Kate think the candidate who was claimed to be successful was?
gap	negative	How fast did Lily think the runner expected not to be?
gap	positive	How fast did Lily think the runner expected to be?
reduced	negative	How fast did Lily think the runner expected not to be winning was?
reduced	positive	How fast did Lily think the runner expected to be winning was?
relative	negative	How fast did Lily think the runner who was expected not to be winning was?
relative	positive	How fast did Lily think the runner who was expected to be winning was?
gap	negative	How calm did Molly say the mother claimed not to be?
gap	positive	How calm did Molly say the mother claimed to be?
reduced	negative	How calm did Molly say the mother claimed not to be furious was?
reduced	positive	How calm did Molly say the mother claimed to be furious was?
relative	negative	How calm did Molly say the mother who was claimed not to be furious was?
relative	positive	How calm did Molly say the mother who was claimed to be furious was?
gap	negative	How strong did John think the boy expected not to be?
gap	positive	How strong did John think the boy expected to be?
reduced	negative	How strong did John think the boy expected not to be sociable was?
reduced	positive	How strong did John think the boy expected to be sociable was?
relative	negative	How strong did John think the boy who was expected not to be sociable was?
relative	positive	How strong did John think the boy who was expected to be sociable was?

Experiment 2 & 3: self-paced reading and eye-tracking

Condition 1	Condition 2	Sentences
reduced	negative	How tall did Mary think the girl hoped not to be famous by her parents was before she went to college?
reduced	positive	How tall did Mary think the girl hoped to be famous by her parents was before she went to college?
RC	positive	How tall did Mary think the girl who was hoped to be famous by her parents was before she went to college?
RC	negative	How tall did Mary think the girl who was hoped not to be famous by her parents was before she went to college?
reduced	negative	How successful did Lily think the attorney hoped not to be wealthy by his family was when she first met him?
reduced	positive	How successful did Lily think the attorney hoped to be wealthy by his family was when she first met him?
RC	positive	How successful did Lily think the attorney who was hoped to be wealthy by his family was when she first met him?
RC	negative	How successful did Lily think the attorney who was hoped not to be wealthy by his family was when she first met him?
reduced	negative	How famous did Fred think the singer hoped not to be thinner by her agent was after she won the prize?
reduced	positive	How famous did Fred think the singer hoped to be thinner by her agent was after she won the prize?
RC	positive	How famous did Fred think the singer who was hoped to be thinner by her agent was after she won the prize?
RC	negative	How famous did Fred think the singer who was hoped not to be thinner by her agent was after she won the prize?
reduced	negative	How objective did John think the editor hoped not to be strict by her colleagues was when selecting presentation materials?
reduced	positive	How objective did John think the editor hoped to be strict by her colleagues was when selecting presentation materials?
RC	positive	How objective did John think the editor who was hoped to be strict by her colleagues was when selecting presentation materials?
RC	negative	How objective did John think the editor who was hoped not to be strict by her colleagues was when selecting presentation materials?
reduced	negative	How frank did Andy think the interviewer hoped not to be constructive by the officer was at the meeting?
reduced	positive	How frank did Andy think the interviewer hoped to be constructive by the officer was at the meeting?
RC	positive	How frank did Andy think the interviewer who was hoped to be constructive by the officer was at the meeting?
RC	negative	How frank did Andy think the interviewer who was hoped not to be constructive by the officer was at the meeting?
reduced	negative	How popular did Mary think the kid wished not to be quiet by his teacher was when he joined the volleyball club?
reduced	positive	How popular did Mary think the kid wished to be quiet by his teacher was when he joined the volleyball club?
RC	positive	How popular did Mary think the kid who was wished to be quiet by his teacher was when he joined the volleyball club?
RC	negative	How popular did Mary think the kid who was wished not to be quiet by his teacher was when he joined the volleyball club?

reduced	negative	How muscular did James think the student wished not to be talkative by his friends was after he started playing football?
reduced	positive	How muscular did James think the student wished to be talkative by his friends was after he started playing football?
RC	positive	How muscular did James think the student who was wished to be talkative by his friends was after he started playing football?
RC	negative	How muscular did James think the student who was wished not to be talkative by his friends was after he started playing football?
reduced	negative	How happy did Chris think the woman wished not to be married by her parents was when she started a new job abroad?
reduced	positive	How happy did Chris think the woman wished to be married by her parents was when she started a new job abroad?
RC	positive	How happy did Chris think the woman who was wished to be married by her parents was when she started a new job abroad?
RC	negative	How happy did Chris think the woman who was wished not to be married by her parents was when she started a new job abroad?
reduced	negative	How brave did Lily think the girl wished not to be popular by her sister was after joining the debate club?
reduced	positive	How brave did Lily think the girl wished to be popular by her sister was after joining the debate club?
RC	positive	How brave did Lily think the girl who was wished to be popular by her sister was after joining the debate club?
RC	negative	How brave did Lily think the girl who was wished not to be popular by her sister was after joining the debate club?
reduced	negative	How adaptable did Robin think the journalist wished not to be incisive by her boss was when she first came to India?
reduced	positive	How adaptable did Robin think the journalist wished to be incisive by her boss was when she first came to India?
RC	positive	How adaptable did Robin think the journalist who was wished to be incisive by her boss was when she first came to India?
RC	negative	How adaptable did Robin think the journalist who was wished not to be incisive by her boss was when she first came to India?
reduced	negative	How slim did John think the doctor wished not to be gentle by his patients was when he graduated from medical school?
reduced	positive	How slim did John think the doctor wished to be gentle by his patients was when he graduated from medical school?
RC	positive	How slim did John think the doctor who was wished to be gentle by his patients was when he graduated from medical school?
RC	negative	How slim did John think the doctor who was wished not to be gentle by his patients was when he graduated from medical school?
reduced	negative	How innocent did Jack think the defendant claimed not to be guilty by the witness was before the jury made their decision?
reduced	positive	How innocent did Jack think the defendant claimed to be guilty by the witness was before the jury made their decision?
RC	positive	How innocent did Jack think the defendant who was claimed to be guilty by the witness was before the jury made their decision?
RC	negative	How innocent did Jack think the defendant who was claimed not to be guilty by the witness was before the jury made their decision?
reduced	negative	How responsible did Brian think the president claimed not to be healthy by his doctor was when he delivered the public speech?

reduced	positive	How responsible did Brian think the president claimed to be healthy by his doctor was when he delivered the public speech?
RC	positive	How responsible did Brian think the president who was claimed to be healthy by his doctor was when he delivered the public speech?
RC	negative	How responsible did Brian think the president who was claimed not to be healthy by his doctor was when he delivered the public speech?
reduced	negative	How heavy did Lily think the player claimed not to be strong by his coach was when he joined the football team?
reduced	positive	How heavy did Lily think the player claimed to be strong by his coach was when he joined the football team?
RC	positive	How heavy did Lily think the player who was claimed to be strong by his coach was when he joined the football team?
RC	negative	How heavy did Lily think the player who was claimed not to be strong by his coach was when he joined the football team?
reduced	negative	How poor did Jason think the child claimed not to be intelligent by her parents was when she started school?
reduced	positive	How poor did Jason think the child claimed to be intelligent by her parents was when she started school?
RC	positive	How poor did Jason think the child who was claimed to be intelligent by her parents was when she started school?
RC	negative	How poor did Jason think the child who was claimed not to be intelligent by her parents was when she started school?
reduced	negative	How calm did Molly say the mother claimed not to be furious by her husband was when she found out her son was expelled from school?
reduced	positive	How calm did Molly say the mother claimed to be furious by her husband was when she found out her son was expelled from school?
RC	positive	How calm did Molly say the mother who was claimed to be furious by her husband was when she found out her son was expelled from school?
RC	negative	How calm did Molly say the mother who was claimed not to be furious by her husband was when she found out her son was expelled from school?
reduced	negative	How nervous did Kate think the candidate claimed not to be successful by the media was before he went on the stage?
reduced	positive	How nervous did Kate think the candidate claimed to be successful by the media was before he went on the stage?
RC	positive	How nervous did Kate think the candidate who was claimed to be successful by the media was before he went on the stage?
RC	negative	How nervous did Kate think the candidate who was claimed not to be successful by the media was before he went on the stage?
reduced	negative	How fast did Lily think the runner expected not to be winning by the audience was for the final round of the race?
reduced	positive	How fast did Lily think the runner expected to be winning by the audience was for the final round of the race?
RC	positive	How fast did Lily think the runner who was expected to be winning by the audience was for the final round of the race?
RC	negative	How fast did Lily think the runner who was expected not to be winning by the audience was for the final round of the race?
reduced	negative	How strong did John think the boy expected not to be sociable by his coach was before the training camp started?
reduced	positive	How strong did John think the boy expected to be sociable by his coach was before the training camp started?

RC	positive	How strong did John think the boy who was expected to be sociable by his coach was before the training camp started?
RC	negative	How strong did John think the boy who was expected not to be sociable by his coach was before the training camp started?
reduced	negative	How convincing did Judy think the officer expected not to be present in the conference was in front of aggressive reporters?
reduced	positive	How convincing did Judy think the officer expected to be present in the conference was in front of aggressive reporters?
RC	positive	How convincing did Judy think the officer who was expected to be present in the conference was in front of aggressive reporters?
RC	negative	How convincing did Judy think the officer who was expected not to be present in the conference was in front of aggressive reporters?
reduced	negative	How smart did James think the child expected not to be polite by his classmates was when he started teaching him?
reduced	positive	How smart did James think the child expected to be polite by his classmates was when he started teaching him?
RC	positive	How smart did James think the child who was expected to be polite by his classmates was when he started teaching him?
RC	negative	How smart did James think the child who was expected not to be polite by his classmates was when he started teaching him?
reduced	negative	How positive did Amy think the interviewee expected not to be hard-working by his interviewers was after the final round of interview?
reduced	positive	How positive did Amy think the interviewee expected to be hard-working by his interviewers was after the final round of interview?
RC	positive	How positive did Amy think the interviewee who was expected to be hard-working by his interviewers was after the final round of interview?
RC	negative	How positive did Amy think the interviewee who was expected not to be hard-working by his interviewers was after the final round of interview?
reduced	negative	How dependable did Adam think the manager expected not to be bossy by his colleagues was in dealing with the new project?
reduced	positive	How dependable did Adam think the manager expected to be bossy by his colleagues was in dealing with the new project?
RC	positive	How dependable did Adam think the manager who was expected to be bossy by his colleagues was in dealing with the new project?
RC	negative	How dependable did Adam think the manager who was expected not to be bossy by his colleagues was in dealing with the new project?
reduced	negative	How pessimistic did Kate think the lawyer expected not to be patient by his clients was about the testimony of the witness?
reduced	positive	How pessimistic did Kate think the lawyer expected to be patient by his clients was about the testimony of the witness?
RC	positive	How pessimistic did Kate think the lawyer who was expected to be patient by his clients was about the testimony of the witness?
RC	negative	How pessimistic did Kate think the lawyer who was expected not to be patient by his clients was about the testimony of the witness?

Article 1

Predicting Coarse-grained Semantic Features in Language Comprehension:

Evidence from ERP Representational Similarity Analysis and Chinese Classifier[1]

Zirui Huang[1,3], Chen Feng[1,2], Qingqing Qu[1,2]

[1]Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
[2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China
[3]Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, United Kingdom

Word count: 5,327 (main text, excl. references)

Qingqing Qu
Key Laboratory of Behavioral Science
Institute of Psychology
Chinese Academy of Sciences, Beijing, China
16 Lincui Road, Chaoyang District, Beijing, China 100101
China
Tel: +86-10-64888629
Fax: +86-10-64872010
Email: quqq@psych.ac.cn

---

[1] This chapter is based on: Huang, Z., Feng, C., & Qu, Q. (2023). Predicting coarse-grained semantic features in language comprehension: evidence from ERP representational similarity analysis and Chinese classifier. *Cerebral Cortex, 33*(13), 8312–8320. https://doi.org/10.1093/cercor/bhad116
The original article is published Open Access under the Creative Commons Attribution License (CC BY 4.0).
Authorship and contribution for this chapter are declared in the Authorship Statement submitted with the thesis.

Abstract

Existing studies demonstrate that comprehenders can predict semantic information during language comprehension. Most evidence comes from highly constraining context, in which a specific word is likely to be predicted. One question that has been investigated less is whether prediction can occur when prior context is less constraining for predicting specific words. Here, we aim to address this issue by examining the prediction of animacy features in low-constraining context, using electroencephalography (EEG), in combination with representational similarity analysis (RSA). In Chinese, a classifier follows a numeral and precedes a noun, and classifiers constrain animacy features of upcoming nouns. In the task, native Chinese Mandarin speakers were presented with either animate-constraining or inanimate-constraining classifiers followed by congruent or incongruent nouns. EEG amplitude analysis revealed an N400 effect for incongruent conditions, reflecting the difficulty of semantic integration when an incompatible noun is encountered. Critically, we quantified the similarity between patterns of neural activity following the classifiers. RSA results revealed that the similarity between patterns of neural activity following animate-constraining classifiers was greater than following inanimate-constraining classifiers, before the presentation of the nouns, reflecting pre-activation of animacy features of nouns. These findings provide evidence for the prediction of coarse-grained semantic feature of upcoming words.

Keywords: Semantic prediction; Pre-activation of semantic features; Chinese classifier; EEG; Representational similarity analysis

Introduction

Probabilistic prediction is hypothesized to be an important computational principle underlying language comprehension (Federmeier, 2007; Huettig, 2015; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018; Pulvermüller & Grisoni 2020). A growing number of studies have demonstrated that people are able to make linguistic predictions at various representational levels including semantic (Altmann, & Kamide, 1999; Federmeier & Kutas, 1999; Lau et al. 2013; Wang et al., 2018; 2020), phonological (DeLong et al. 2005; Li et al., 2022; Vissers et al. 2006), written form (Laszlo & Federmeier 2009; Kim & Lai, 2012), and morphosyntactic (Dikker et al., 2009, 2010; Van Berkum et al. 2005) features of words. One question that has been investigated less is whether prediction can occur when prior context is less constraining for predicting specific words but can elicit the prediction of more general information. Most evidence comes from sentences with highly constraining context, in which a specific word is likely to be predicted. Here, we aim to address this issue by examining the prediction of semantic features in low-constraining context, using electroencephalography (EEG), in combination with representational similarity analysis (RSA).

A large group of studies support (not necessarily demonstrate) that people are able to predict semantic information of upcoming words. In a classic event-related potential (ERP) study by Federmeier and Kutas (1999), participants were instructed to read pairs of sentences (e.g., "They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of…"). The following word was either a predictable word (palms), an unpredictable word, but from the same semantic category (pines), or an unpredictable word from a different semantic category (tulips). They found that unpredictable words from the same category (pines)

elicited smaller N400s compared to the unpredictable word from an unrelated category (tulips). The N400 effect varied as a function of the cloze probability of words, but not the plausibility of sentences, suggesting the effect reflects prediction-related pre-activation rather than integration of the target word with preceding context. Until recently, pre-stimulus predictive brain activity (i.e. prediction potential) was discovered (Grisoni et al. 2016, 2017, 2021; León-Cabrera et al. 2017; see Pulvermüller & Grisoni 2020 for a review). These studies have reported that high-constraining context induced anticipatory brain activity preceding the expected words, and the size of the anticipatory brain activity is correlated with the predictability of the expected word (see Grisoni et al. 2021). Converging evidence for semantic prediction comes from anticipatory eye movements toward objects before the expressions that refer to these objects. For instance, in a visual world task, Kamide et al. (2003) manipulated the agent of an action, and reported that when the preceding context was “the man will ride …”, listeners predictively looked at “motorbike” more than “carousel” before the presentation of the following nouns. The findings suggest that comprehenders use world knowledge about plausible actions by agents to predict semantic information of plausible actions.

Most work concerns the prediction of a specific word in a highly constraining context (Wicha et al., 2004; Ito et al., 2018; Wang et al., 2018; Li et al., 2022). It is less clear whether prediction can occur when prior context is less constraining for predicting specific words but can elicit the prediction of more general information that characterize a group of multiple words. Investigating whether comprehenders are only able to predict specific words at the fine-grained level or are also able to predict upcoming information at a coarse-grained level is theoretically important, because contexts that predict specific words (e.g., “I like coffee with sugar and ...”)

are not frequent, and most words in natural language are not highly predictable on the basis of prior contexts or knowledge (e.g., I'd like to drink...) (Luke & Christianson, 2016).

In the present study, we adopted the constraints of the classifier system in Mandarin Chinese, to test whether comprehenders can utilize the semantic constraints of classifiers to predict animacy feature of upcoming nouns. Chinese is a numeral classifier language, in which a classifier is obligatory between a numeral and a noun when a noun is modified by a numeral (e.g., one, two), a demonstrative (e.g., this, that), or a quantifier (e.g., a few) (Li & Thompson, 1989), which contrasts to English where a noun immediately follows a numeral (e.g., one person). In other words, a classifier follows a numeral and precedes a noun in Chinese (e.g., 一个人, one “ge4” person). Chinese speakers need to select a proper classifier from 174 classifiers (c.f., Huang et al, 1997) based on semantic features of the accompanying noun. Therefore, a classifier can constrain semantic features of its following noun, such as animacy, shape, size, etc. Although the mapping between a classifier and a noun can be to some degree arbitrary so that a classifier can modify different nouns, and a noun can be preceded by various classifiers, some Chinese classifiers are unambiguous in specifying animacy features of nouns, therefore constraining the animacy of following nouns. Some classifiers go with inanimate nouns only (e.g., 朵, /duo/, classifying flowers) whereas some classifiers modify animate nouns (e.g., 名, /ming/, classifying human noun, such as teacher, doctor etc.).

The classifier-noun association, combined with the violation paradigm where sentences are presented with classifier-noun mismatch errors, is commonly used to investigate semantic integration processes during Chinese sentence comprehension. A typical finding is that the classifier-noun mismatch elicits N400 effects on the noun, with larger negativity for mismatch

trials, which reflects the difficulty of integrating the lexical semantics into the representation at the higher level (e.g., Zhou et al., 2010; Zhang et al., 2012). The regularity of the classifier-noun agreement, combined with ERPs, can provide efficient means to examine the effects of prediction during language comprehension. In one of few studies concerning semantic prediction in Chinese, Kwon et al. (2017) manipulated whether a classifier embedded in a sentence matched or mismatched an upcoming expected noun. Kwon et al. replicated the well-established N400 effect, with enhanced N400 amplitude to unexpected nouns. The N400 was also evident as early as the preceding classifier, suggesting the pre-activation of semantic features of nouns. However, the effect on classifiers may reflect integration of the classifier with the preceding context, rather than preactivation of the noun (also see Szewczyk & Schriefers, 2013). It is very difficult (if not impossible) to distinguish prediction from integration, and demonstrate evidence that is compatible with prediction but not integration, by analyzing ERPs after encountering a target word, using the traditional ERP amplitude analysis.

In addition to ERP amplitude analysis, predictive behavior in language comprehension has recently been investigated using (Kriegeskorte et al., 2008), i.e., analyzing similarity among patterns of neural activity, before or after encountering the target word. The basic assumption of RSA is that similarities between items can elicit similarities in patterns of brain activity. Wang et al. (2018) adopted RSA combined with magnetoencephalography (MEG) to examine whether comprehenders can predict specific words when they read highly constraining sentences. Wang et al. quantified the similarity between patterns of brain activity, and found that the patterns of neural activity were more similar when the same words were predicted than when different words were predicted, critically before the onset of the predicted words. The results

demonstrate that the prediction of specific words is associated with unique patterns of neural activity. In another study using RSA, Hubbard and Federmeier (2021) reanalyzed the previous EEG sentence reading study (Federmeier et al., 2007), and compared EEG activity similarity patterns elicited by sentence final words to patterns from the preceding words of the sentence. The logic is that if pre-final words elicit the pre-activation of features of the final word in constraining sentences, then some aspects of the neural representation of the final word should appear during the processing of the pre-final word, thus producing greater similarity between pre-final and final words. Measuring pattern similarity of the sentence final word and words prior to the pre-final word revealed that neural similarity with the final word was increased following the processing of only the pre-final word. The effect was not observed in earlier words and the increase was modulated by both final word expectancy and sentence constraint. These findings demonstrate a precisely timed semantic prediction.

In more recent work, Wang et al. (2020) used RSA to detect neural patterns for the prediction of animacy features of upcoming nouns when sentence contexts only constrain coarse-grained semantic animacy feature as opposed to a specific word. Semantic animacy of nouns were constrained by verbs (e.g., “caution” constrains for animate nouns; “unfold” constrains for inanimate nouns). RSA revealed that before the onset of nouns, patterns of neural activity were more similar following animate-constraining verbs than following inanimate-constraining verbs, providing evidence for the prediction of coarse-grained semantic features that goes beyond the prediction of individual words. One caveat with this type of design is that neural similarity may not reflect the similarity in the meanings of the predicted words but rather the similarity of the meanings of the preceding context, which is bound to correlate with the predicted target word.

In the present study, we reduce prior context and directly compare the EEG patterns after the presentation of classifiers without any additional constraining inputs. The goal is thus to investigate whether comprehenders can use the constraints provided by classifiers to predict semantic features associated with the animacy of upcoming nouns. We focus on the similarity between patterns of brain activity following the classifiers until just before the presentation of the nouns, although we analyze the similarity value after the presentation of nouns. Because animate entities share more strongly associative semantic features than inanimate entities, which have more distinctive features (McRae et al., 1997; Daniele Zannino et al., 2006), the brain activity patterns between word pairs should be more similar among animate nouns than inanimate nouns during the presentation of nouns. Critically, the question is whether the difference in similarity occurs before the presentation of nouns. If comprehenders predict the animacy of upcoming nouns, the similarity between patterns of neural activity should be greater following animate-constraining classifiers than following inanimate-constraining classifiers.

Method

Participants.

In total, 29 native speakers of Mandarin Chinese who were resident in Beijing participated in the ERP experiment. Sample size was determined by recent EEG/MEG RSA studies of language prediction (Hubbard and Federmeier, 2021; Wang et al. 2018). EEG data from four participants was excluded for data analysis due to a high percentage of rejected trials (more than 50%) or an empty data set for at least one classifier after data preprocessing, and thus 25 participants were included for ERP and RSA analyses. All participants had normal or correct-to-normal vision and no history of language disorders. They were given informed consent and paid about ~$20. The

study was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences.

Materials and design.

A total of 12 classifiers were used, and each classifier was combined with 10 nouns to form 120 semantically plausible classifier-noun phrases. Among the 12 classifiers, six of them were animate-constraining classifiers that can only modify animate nouns. Of these, three were human-modifying classifiers and three animal-modifying classifiers; the other six classifiers were inanimate-constraining classifiers including three natural-object-modifying classifiers and three artifact-product-modifying classifiers. We have assessed the close probability of the classifiers using the cloze test. A group of 25 Chinese native speakers who did not take part in the EEG experiment were presented with incomplete phrase (i.e. Numeral + Classifier) and were asked to complete fragments with the most likely ending. Classifier constraint was measured as the proportion of participants who gave the most frequent word. Overall, the cloze probability of the classifiers was low (25.7% ± 11.3%), indicating that classifiers used in our study do not produce specific word predictions. Moreover, cloze probability was matched between animate-constraining classifiers (animate: 26.0% ± 7.0%) and inanimate-constraining classifiers (inanimate: 25.3% ± 1.6%), $P = 0.934$. Besides, these classifiers can highly constrain animacy of nouns as expected (probability of animate nouns following animate-constraining classifiers: 93.3%, probability of inanimate nouns following inanimate-constraining classifiers: 97.3%).

Each of 12 classifiers was recombined with incongruent nouns to form 120 semantically implausible classifier-noun phrases, including 60 semantically incongruent but animacy-

matched trials (i.e., Incongruent, Animacy-Match), in which classifiers were re-paired with nouns from different subgroups (e.g., human classifiers were re-paired with animal nouns; artifact-product classifiers were paired with natural object nouns) and 60 semantically incongruent AND animacy-mismatched trials (i.e., Incongruent, Animacy-mismatch), in which classifiers were paired with nouns from different animacy groups (e.g., human classifiers were recombined with natural object nouns; artifact product classifiers were recombined with animal nouns), see Table 1 for examples. The phrases were more implausible in the *Incongruent, Animacy-mismatch* compared to the *Incongruent, Animacy-Match* condition, due to the additional mismatch in animacy. That is, there was a graded difference in the degree of meaning congruence of the nouns and their classifiers across the three conditions. In total, 240 combinations of "classifier + noun" were used in the study. Each participant was presented with three blocks of 80 trials with the 40 plausible and 40 implausible classifier-noun combinations in a randomized order.

Conditions	Animate-constraining classifier	Inanimate-constraining classifier
Congruent	一位 教授 one CL_{human} professor	一本 杂志 one CL_{book} magazine
Incongruent, Animacy-Match	一位 骏马 one CL_{human} horse	一本 椅子 one CL_{book} chair
Incongruent, Animacy-Mismatch	一位 杂志 one CL_{human} magazine	一本 教授 one CL_{book} professor

Table 1. Experimental conditions and examples of stimuli

Quantifying the semantic and lexical similarity of the animate- vs. inanimate- constraining classifiers.

Our aim was to measure the neural activity elicited by the pre-activation of animacy feature of the upcoming noun. Therefore, we focused on activity following the onset of the classifier until just before the onset of subsequent nouns. To make sure that any difference in the representational similarity of ERPs before the onset of nouns indeed reflected the pre-activation of the upcoming nouns rather than similarity associated within the animate- vs. inanimate-constraining classifiers, we verified that the two groups of classifiers matched on several key properties relevant to visual and linguistic processing. These properties included semantic similarity, visual complexity and word frequency. To quantify semantic similarity between word pairs among the animate- vs. inanimate- constraining classifiers, we used HowNet, an online database that provide calculations of inter-conceptual and inter-attribute relationships of Chinese lexicons (Dong et al., 2010). Semantic similarity values for all possible classifier pairs were measured via a path-based approach by Wu and Palmer (1994). These pairwise Wu-Palmer semantic similarity values in a 12 by 12 matrix are presented in Figure 1A. The mean semantic similarity values for the animate- vs. the inanimate- classifiers showed no difference (t = 0.15, p = 0.88). We also verified that other aspects of lexical similarity (i.e., visual complexity, and word frequency) associated with the animate- vs. inanimate- constraining classifiers were matched. In order to do that, we extracted visual complexity of Chinese characters of classifiers (the number of strokes) and word frequency (based on the Chinese Linguistic Data Consortium norms,2013). The similarity values of visual complexity and word frequency were measured by the absolute difference for each possible pair of classifiers. Statistical analysis showed that visual

complexity (t = -0.2, p = 0.84) and word frequency (t = 1.18, p = 0.26) were matched between the animate- and the inanimate- classifiers.

Quantifying the semantic and lexical similarity of the animate vs. inanimate nouns.

The experimental hypothesis rested on the assumption that animate nouns constrained for by their classifiers would be more semantically similar to each other than inanimate nouns. We used the same approach as described above to verify this assumption: We calculated the semantic similarity values between all possible pairs of nouns within the animate-constraining and inanimate-constraining conditions. As shown in Figure 1B, the mean semantic similarity within the animate set was indeed greater than that within the inanimate set (t = 2.24, p = 0.03). Moreover, it is important to confirm that any differences in neural similarity produced by predicted animate and inanimate nouns were not generated by differences in similarity of lexical properties. Statistical analysis showed that visual complexity and word frequency were matched between animate vs. inanimate nouns (visual complexity: t = 0.48, p = 0.63; word frequency: t = 1.14, p = 0.26).

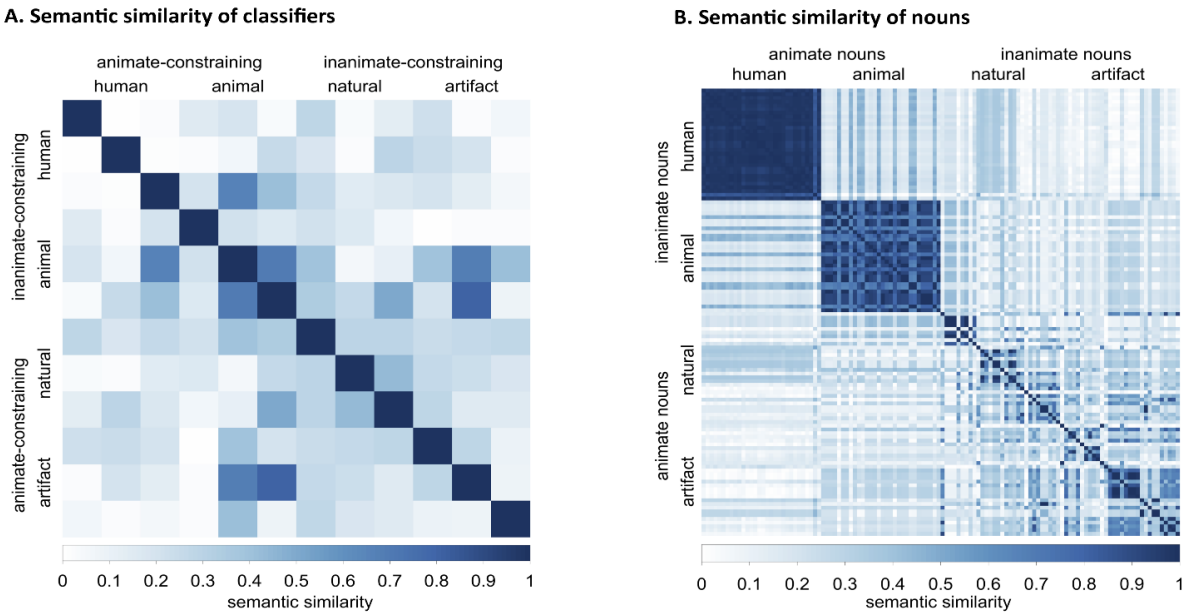


Figure 1. Pairwise Wu and Palmer similarity values for classifiers (A) and nouns (B). Similarity values are ranging from 0 to 1, with 0 indicating no similarity and 1 indicating identical semantics. A shows a 12 by 12 symmetric semantic similarity matrix with 6 animate constraining classifiers (3 human classifiers and 3 animal classifiers) and 6 inanimate classifiers (3 natural object classifiers and 3 artifact classifiers). B shows a 120 by 120 semantic similarity matrix of nouns that are semantically matched with each classifier (each classifier is paired with 10 nouns), which confirmed the animacy constraint of classifiers.

Procedure.

The experiment was conducted using E-Prime software. Participants were first instructed that they would see the classifier-noun phrases presented on the computer screen and their task was to judge whether the classifier-noun phrase was semantically plausible or not by pressing a designated keyboard button. Each trial started with an asterisk signal for 500 ms, followed by a presentation of a numeral and a classifier for 1,000 ms, a blank screen for 1,000 ms and a noun for 1,000ms. An interval of 1,000 ms was inserted between trials. Participants received four practice trials before seeing 240 experimental trials presented in three experimental blocks and separated by a short break. The experimental task was 25 minutes and the entire experiment lasted approximately 90 minutes.

EEG recordings and analysis.

EEG signals were collected from 64 electrodes secured in an elastic cap and recorded in Neuroscan software. The vertical electrooculogram (VEOG) was captured via two electrodes above and below the left eye. The horizontal electrooculogram (HEOG) was recorded via two electrodes on the left and right external cantus. The left mastoid electrode was used as a reference. The EEG data were re-referenced to the average of both mastoids. All electrode impedances were kept below 5 kΩ. EEG signals were amplified with a band-pass filter between 0.05 and 70 Hz with a sampling rate of 1000 Hz). The EEGLAB was used for the preprocessing of EEG data. Raw data were down-sampled to 500 Hz and filtered with a high-pass cutoff point of 0.1 Hz and a low-pass cutoff point of 30 Hz. Independent Component Analysis (ICA) using the infomax ICA algorithm of Bell & Sejnowski (1995) was performed to remove vertical or horizontal eye movements, channel noise, muscle artifacts, and 1–3 ICs were removed per participant. Epochs containing amplitudes exceeding ±70 μV were rejected (~4.6% of all epochs). Trials with incorrect responses were excluded from analyses (4.7%). The remaining epochs for ERP analysis were on average 110 trials per condition (animate-constraining vs. inanimate-constraining). The EEG was segmented into 3,500 ms epochs that included a 500 ms pre-stimulus display as a baseline, a 1,000 ms classifier display, 1,000 ms interval, and 1,000 ms noun display. Epochs were baseline corrected and the averaged signal in a baseline period was subtracted from the remaining ERPs.

Mean amplitude analyses were performed for pre-noun and after-noun time windows separately. First, mean amplitude analyses for pre-noun time window were performed to examine pre-stimulus predictive brain activity, a negative shift of brain potentials before the

onset of the predictable target (i.e. “prediction potential”). To this end, neural activity for animate-constraining versus inanimate-constraining trials for the time window (−200 ms, 0 ms) was analyzed with a baseline correction computed across (−300 ms, −200 ms) before noun onset. We exacted the mean ERP amplitudes at fronto-central electrodes (FC1, FCZ, FC2, C1, CZ, C2, CP1, CPZ, CP2), where the prediction potential is observed (e.g. Grisoni et al. 2017). We investigated whether animate- and inanimate-constraining classifiers can elicit significant negative deflection, by testing pre-noun ERPs against zero, and whether there is difference between animate- versus inanimate-constraining classifiers by directly comparing pre-noun ERPs of two conditions. Second, mean amplitude analyses for after-noun time window were performed to investigate the semantic incongruent effect evoked by the incongruency between classifiers and nouns. Based on previous findings (as reviewed in the Introduction), we expected a N400 effect evoked by incongruent nouns with classifiers. Nine regions of interest (ROIs) were defined to examine the distribution of effects on the scalp. The ROIs include the left-anterior area (F3, F5, FC3), the middle-anterior area (FZ), the right-anterior area (F4, F6, FC4), the left-middle area (C3, C5), the middle–middle area (CZ), the right-middle area (C4, C6), the left-posterior area (P3, P5), the middle-posterior area (PZ), and the right-posterior area (P4, P6). The time window of 300–500 ms was chosen based on visual inspection for N400 effects. Mean amplitudes on this time window were entered into a 2 × 3 × 9 repeated measures analysis of variance (ANOVA) with the factors classifier-type (animate-constraining classifier/inanimate-constraining classifier), congruency condition (Congruent/Incongruent, Animacy-Match/Incongruent, Animacy-Mismatch), and ROIs.

RSA analyses were also performed for pre-noun and after- noun time windows separately. First, RSA analyses for pre-noun time window were performed to examine pre-activation of animacy features. We computed an EEG vector across all channels (62 electrodes) and time points by averaging the data for each classifier. For each individual participant, we calculated Pearson's r values to determine the similarity values between the patterns of neural activity following all possible pairs of the animate-constraining classifiers (6∗5/2 = 15 pairs) and those of the inanimate-constraining classifiers (6∗5/2 = 15 pairs) at each time point from the onset of the classifiers until the onset of nouns. We then averaged these pairwise correlation r values to yield averaged similarity values at each time point for each participant for animate- and inanimate-constraining conditions. We then conducted statistical analyses to examine whether the neural similarity values for the animate- versus inanimate- constraining conditions are different before the presentation of nouns. To visualize any differences between the two conditions, we averaged these similarity values across all participants at each consecutive time point for animate- and inanimate-constraining conditions. This generated a grand average similarity over time for each condition (see Fig. 2A). Second, RSA analyses for after-noun time window were performed to test whether the neural similarity for the animate nouns is greater than that for the inanimate nouns, as assumed. Each of 12 classifiers was paired with 10 nouns. For each individual participant, we averaged the data of 10 nouns for each classifier, and calculated Pearson's r values to determine the similarity values between the patterns of neural activity following all possible pairs of the animate condition and those of the inanimate condition at each time point during the presentation of nouns. We then averaged these pairwise correlation r values to yield averaged similarity values at each time point for each participant for animate

and inanimate nouns, and then conducted statistical analyses to examine the difference. For statistical analyses, we conducted a paired t-test at each time point across the entire epoch, and critical time windows were identified based on paired t-test results that exceeded a preset uncorrected P-value threshold of 0.05. For each critical time window, the observed summed t-value was determined by summing the individual t-values at each time point testing within the time window. We used a nonparametric permutation procedure to protect against problems associated with multiple time points (Maris and Oostenveld 2007). We randomly shuffled the condition labels for each participant, and within each critical time window, we performed t-test at each time point and summed individual t-values. This procedure was repeated for 1,000 times and formed H0 distribution of summed t-values. The observed summed t-value falls outside the 95% range is considered to be significant. Separate permutation tests were performed to examine differences for pre-noun and after-noun time windows.

Results

Behavioral data.

Participants were asked to respond to incongruent phrases, so behavioral data were only available for incongruent phrases. The overall mean response latency for incongruent classifier-noun phrases was 663 ms (standard deviation SD = 195 ms). Response latencies for the Incongruent, Animacy-Mismatch condition (652 ms) were faster than for the Incongruent, Animacy- Match condition (674 ms), suggesting that judgments were made faster for trials with additional animacy-mismatch. A linear mixed-effect model analyses confirmed this observation, showing a significant difference between both conditions (P = 0.027, df = 118).

EEG data.

We first compared the ERPs of the two conditions to determine if the larger spatial similarities following the animate- versus inanimate-constraining classifiers could be explained by differences in the ERPs evoked by these classifiers. As expected, the ERPs evoked by animate- versus inanimate-constraining classi- fiers were similar, and a cluster-based permutation test over the entire epoch failed to reveal a significant ERP effect (P = 0.359).

Mean amplitude analyses for pre-noun time window demon- strate a negative shift of brain potentials before the onset of nouns (i.e. “prediction potential,” see Fig. 2B). In the last 200 ms before noun onset, both animate- and inanimate-constraining classifiers elicited significant negative-going potentials, as documented by t-tests against zero (ps < 0.001). Animate- and inanimate—constraining classifiers did not elicit difference, as expected (P = 0.43). Mean amplitude analyses for after-noun time window revealed the semantic incongruent effect evoked by the incongruency between classifiers and nouns. Fig. 2C and D show grand average ERPs for the three congruency conditions at a rep- resentative electrode (Cz) under animate classifiers (Fig. 2C) and inanimate classifiers (Fig. 2D). ANOVAs on the mean amplitude of nouns with the factors classifier type, congruency condition, and ROIs revealed main effects of congruency condition (F = 38.26, P < 0.001) and classifier types (F = 24.00, P = 0.03), an interaction between the three factors (F = 127.66, P < 0.001), and a marginally significant interaction between classifier type and congruency condition (F = 48.00, P = 0.07). The effect of congruency con- dition revealed that compared with the congruent condition, incongruent conditions elicited a larger negativity in the 300– 500 ms time window with a broad distribution in all nine ROIs (P-values false discovery rate corrected): left-anterior (t = 6.03, P < 0.001), left-

middle (t = 6.47, P < 0.001), left-posterior (t = 3.98, P < 0.001), middle-anterior (t = 5.89, P< 0.001), middle-middle (t = 6.14, P < 0.001), middle-posterior (t = 3.69, P = 0.001), right-anterior (t = 7.03, P < 0.001), right-middle (t = 6.61, P < 0.001), and right-posterior (t = 4.74, P < 0.001). Following the interactions, separate analyses were conducted for each classifier type. For animate classifiers, there was a graded effect with the largest negativity in the Incongruent, Animacy-Mismatch condition, compared to the Incongruent, Animacy-Match condition, due to the additional mismatch in animacy. Pairwise comparisons for each ROI revealed the additional mismatch effect in animacy with a broad distribution, i.e. left-posterior (t = 2.55, P = 0.026), middle- middle (t = 2.59, P = 0.016), middle-posterior (t = 2.38, P = 0.038), right-middle (t = 2.24, P = 0.035), and right-posterior (t = 3.11, P = 0.007). Contrary to animate classifiers, for inanimate clas- sifiers the additional mismatch effect from animacy disappeared in all nine ROIs (ts < 1.52; ps > 0.14).

Grand average similarity waveforms over time are displayed in Fig. 2A for animate- versus inanimate-constraining classifiers. Statistical analyses revealed that during the presentation of classifiers (−2,000 ms to –1,000 ms), there was no greater neural similarity for animate-constraining classifiers, relative to inanimate-constraining classifiers. The neural similarity within the animate-constraining classifier group was greater than that of the inanimate-constraining condition, from 240 ms before the onset of noun. Statistical analyses confirmed that the greater neural similarity within animate-constraining classifier group was observed before noun onset, relative to neural similarity within the inanimate-constraining condition (−240– 0 ms; P = 0.036), reflecting the prediction of animacy features associated with the upcoming words. RSA analyses for after-noun time window revealed that the patterns of brain activity were more similar among

animate nouns than among inanimate nouns throughout the whole time window of noun presentation (P = 0.04). To test for the possibility that word frequency would explain the RSA effects, we used linear mixed-effects models to predict the neural similarity value in the predictive time window with type of animacy constraint and word frequency. We measured frequency similarity by taking the absolute value of the difference of word frequency between classifiers, and included this measure in the model. The maximal model included random intercepts for participants and items, and participant random slopes for type of animacy constraint and frequency similarity. Because type of animacy constraint were manipulated between items, item slope adjustments were not specified for the factor. In cases in which the maximal model failed to converge, we sequentially simplified the random effects until convergence was achieved. The mixed- effects model analysis was performed using the "lme4" package as implemented in R. Significance of fixed effects of the model was assessed with the anova() function from package lmerTest, using the Satterthwaite method of approximation for degrees of freedom. Results revealed that the classifier type (animate vs. inanimate) remained a significant effect (F = 3.20, P = 0.048), even with word frequency included. The effect of frequency failed to reach significance (F < 1).

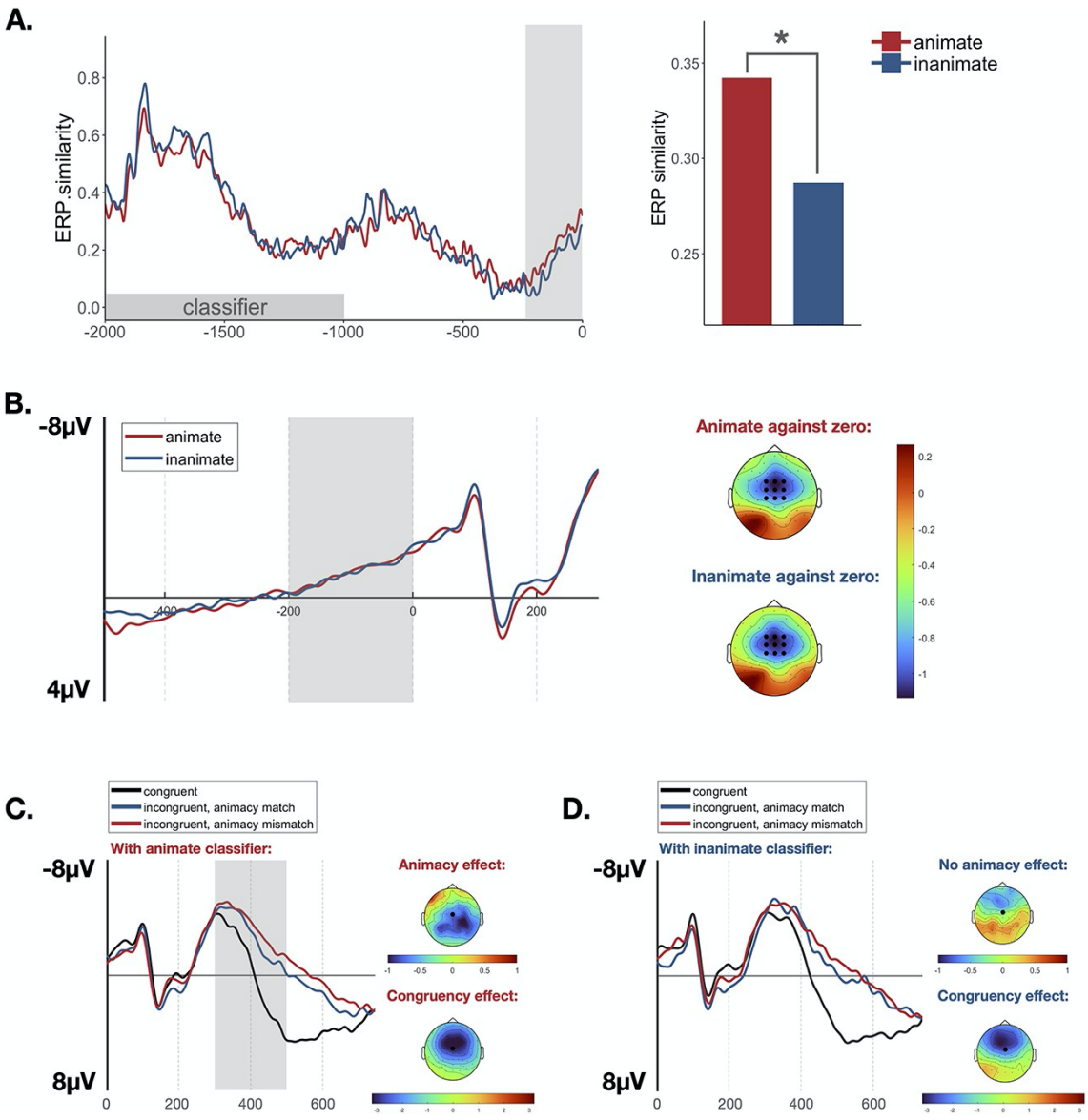


Figure. 2 RSA and ERP results. (A) Pre-noun RSA results. It shows similarity values of animate classifier (red line) and inanimate classifier (blue line) before the onset of nouns. The similarity between patterns of neural activity following animate-constraining classifiers was greater than following inanimate-constraining classifiers ~240 ms before the onset of nouns. The bar chart presents differences in similarity values from −240 ms to 0 ms between animate- (red bar) versus inanimate- (blue bar) constraining classifiers. (B) Pre-noun ERP waveforms in the animate- (red line) and inanimate- (blue line) constraining classifiers as the average of the fronto-central electrodes (FC1, FCZ, FC2, C1, CZ, C2, CP1, CPZ, CP2). In the last 200 ms before noun onset, both animate- and inanimate-constraining classifiers elicited significant negative-going potentials. (C) and (D) After-noun N400 effects elicited by classifier-noun incongruency at Cz. (C) Shows for the animate-constraining classifiers, additional animacy-mismatch elicits larger N400 in the shaded time window (300–500 ms), revealed by larger N400 by the incongruent and animacy-mismatch condition (red line) than the incongruent and animacy-match condition (blue line). (D) Shows no additional animacy-mismatch effect for inanimate-constraining classifiers (no significant difference between red line and blue line).

Discussion

In the present study, RSA on EEG data was conducted to investigate whether animacy features would be pre-activated without a highly constraining context. Chinese classifiers were varied in animacy constraint (animal-constraining and inanimate- constraining) and were followed by Congruent, Incongruent but animacy-matched, or Incongruent but animacy-mismatched nouns. Behavioral response latencies showed a faster response (22 ms) for the Incongruent, Animacy-Mismatch condition than the Incongruent, Animacy-Matched condition, due to additional animacy violation. N400 effect was observed for incongruent conditions, with larger negativity for an additional mismatch in animacy (only for animate-constraining classifiers). Both animate- and inanimate-constraining classifiers elicited a negative shift of brain potentials before the onset of nouns, which is likely to reflect a prediction potential. Of greatest relevance to the study was the RSA result, i.e. greater similarity among patterns of neural activity following the animate-constraining than following the inanimate-constraining classifiers. The similarity effects were significant from ∼240 ms before the onset of nouns, which suggests the pre-activation of semantic features of following nouns.

The ERP effects of animacy and semantic violations of nouns

The present study manipulated two types of violation between classifiers and following nouns. For the semantic violation, nouns were semantically incongruent with the preceding classifiers. For the animacy mismatch, there were additional conflict of animacy (the Incongruent, Animacy-Mismatch condition), or no conflict of animacy (the Incongruent, Animacy-Match condition). Semantic violation elicited N400 effects, which is compatible with previous findings, reflecting

difficulty in semantic integration (e.g. Federmeier and Kutas 1999; Zhang et al. 2012; Zhou et al. 2010). Moreover, the additional mismatch in animacy resulted in a further increase in N400 amplitude, suggesting that animacy information of classifiers and nouns is processed during classifier-noun phrases. Interestingly, this additional animacy- mismatch effect only emerged for animate-constraining classifiers, but not for inanimate-constraining classifiers. The null effect of animacy-mismatch for inanimate classifiers is in line with the results from the closely matched counterpart by Zhang et al. (2012) where only inanimate classifiers were included, and no additional animacy-mismatch effect was observed. Our study extended the investigation of animacy-mismatch to animate classifiers and observed the animacy-mismatch effect. The divergence in animacy-mismatch effect between animate versus inanimate classifiers may reflect the possibility that animate classifiers are more constraining for animacy relative to inanimate classifiers.

The classifier-driven prediction of animacy features of upcoming nouns

Recently, a negative-going potential shift starting hundreds of milliseconds prior to predictable stimuli has been highlighted as a neurophysiological index of prediction (e.g. Pulvermüller and Grisoni 2020). The cortical sources underlying the prediction potential reflect specific perceptual and semantic features of anticipated stimuli before predictable stimuli appear (not before unpredictable ones), which suggests its predictive nature (Grisoni et al. 2016, 2017, 2019, 2022). In the present study, animate- and inanimate-constraining classifiers elicited such brain potentials before the onset of nouns, which is likely related to prediction.

More critically, the present study reveals neural similarity effects before the presentation of nouns, which provides more direct evidence for the pre-activation of coarse-grained semantic features that distinguish between upcoming animate and inanimate items. Our results cannot be explained by lexical- semantic processing of the classifiers across conditions because they were matched well in semantic similarity structure and other linguistic properties. In addition, the animacy RSA effect was not found during the presentation of classifiers and was present only after classifiers and before the onset of nouns, which strongly suggests that it was not elicited by processing of classifiers. Moreover, the critical assumption of the present study is that animate nouns share more common semantic features than inanimate nouns, and thus animate nouns are more similar in neural activity patterns than inanimate nouns. It is important to confirm this assumption. In the present study, the mean semantic similarity within the animate nouns was indeed greater than within the inanimate nouns, and correspondingly animate nouns elicited greater neural similarity than inanimate nouns.

As reviewed in Introduction, previous investigations have mainly focused on the pre-activation of specific lexical words. However, in natural language, most contexts cannot constrain a specific lexical item, but likely constrain semantic features. Therefore, it is theoretically important that we verified that comprehenders can predict semantic information of upcoming words beyond individual words. This finding aligns with Wang et al. (2020) in which participants hear three-sentence context, followed by animate-constraining or inanimate-constraining verbs, and show more similar patterns of neural activity following animate-constraining verbs than following inanimate- constraining verbs. One caveat with this type of design is that neural similarity may reflect not the similarity in the meanings of the predicted words but rather the similarity of the

meanings of the preceding context. To minimize the potential possibility, in the present study, we reduced prior context and only used a single classifier to constrain animate or inanimate nouns, which can provide clear evidence for the pre-activation of semantic features.

In sum, we provide neural evidence for the prediction of coarse- grained animacy-related semantic features driven by isolated Chinese classifiers.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32171058 and No. 62061136001), Youth Innovation Promotion Association (Chinese Academy of Sciences), Youth Elite Scientist Sponsorship Program (No. YESS20200138, China Association for Science and Technology), and the Scientific Foundation of Institute of Psychology (No. E2CX3625CX, Chinese Academy of Sciences) to Qingqing Qu.

Authors' Contributions

Z.R.H. and Q. Q. Q designed the study. Z.R.H. and C.F. performed data analyses. Z.R.H. and Q.Q.Q. wrote the paper. Zirui Huang (Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing—original draft, Writing—review & editing), Chen Feng (Formal analysis, Software) and Qingqing Qu (Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing—original draft, Writing—review & editing).

Data Availability

For protection of participants' privacy, the data are available upon request to the authors. Please email the corresponding author for more information.

Notes

Conflict of Interest: The authors declare no conflict of interests.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73,* 247–264.

Caramazza, A., & Shelton, J. R. (1998). Domain- specific knowledge systems in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.

Chinese Linguistic Data Consortium. (2003). 现代汉语通用词表 [Chinese lexicon] (CLDC-LAC-2003-001). Beijing, China: Tsinghua University, State Key Laboratory of Intelligent Technology and Systems, and Chinese Academy of Sciences, Institute of Automation.

Daniele Zannino, G., Perri, R., Pasqualetti, P., Caltagirone, C., & Carlesimo, G. A. (2006). Analysis of the semantic representations of living and nonliving concepts: a normative study. *Cognitive Neuropsychology*, *23*(4), 515-540.

Delong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121.

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition, 110*(3), 293-321.

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science, 21*(5), 629-634.

Dong, Z., Dong, Q., & Hao, C. (2010). Hownet and its computation of meaning. In *Coling 2010: Demonstrations* (pp. 53-56).

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

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

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

Grisoni, L., Dreyer, F. R., Pulvermüller, F. (2016). Somatotopic semantic priming and prediction in the motor system. *Cerebral Cortex, 26,* 2353–2366.

Grisoni, L., Miller, T. M., Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *Journal of Neuroscience, 37,* 4848–4858.

Grisoni, L., Tomasello, R. & Pulvermüller, F. (2021). Correlated brain indexes of semantic prediction and prediction error: Brain localization and category specificity. *Cerebral Cortex, 31,* 1553-1568.

Huang, C., Chen, K., & Lai, C. (Eds.). (1997). 國語日報量詞典*[Mandarin daily dictionary of Chinese classifiers]*. Taipei: Mandarin Daily Press.

Hubbard, R. J., & Federmeier, K. D. (2021). Representational pattern similarity of electrical brain

activity reveals rapid and specific prediction during language comprehension. *Cerebral Cortex*, *31*(9), 4300-4313.

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

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1-11.

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

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience, 24*(5), 1104-1112.

Kriegeskorte N, Mur M, Bandettini P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience,* 4.

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

Kwon, N., Sturt, P., & Liu, P. (2017). Predicting semantic features in Chinese: Evidence from ERPs. *Cognition*, *166*, 433–446.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61,* 326–338.

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

León-Cabrera, P., Rodríguez-Fornells, A., Morís, J. (2017). Electrophysiological correlates of semantic anticipation during speech comprehension. *Neuropsychologia, 99,* 326–334.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A Functional Reference Grammar* (Vol. 3). Univ of California Press.

Li, X., Li, X., & Qu, Q. Q. (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking task in Mandarin Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(5), 531–547.

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

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J*ournal of Neuroscience Methods, 164,* 177– 190.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2),

99–130.

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

Pulvermüller, F., & Grisoni, L. (2020). Semantic prediction in brain and mind. *Trends in Cognitive Sciences, 24,* 781–784.

Van Berkum, J. J. A., Brown, C. M., Hagoort, P., Zwitserlood, P., & Kooijman, V. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 443–467.

Vissers, C. T., Chwilla, D. J. & Kolk, H. H. (2006) Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research, 1106,* 150–163.

Wang, L., Kuperberg, G., & Jensen, O. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *eLife*, *7*, e39061.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: Evidence from MEG and EEG representational similarity analysis. *Journal of Neuroscience*, *40*(16), 3278–3291.

Wicha, N. Y., Moreno, E. M., & 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.

Wu, Z., & Palmer, M. (1994.) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp133–138. Stroudsburg, PA: Association for Computational Linguistic.

Zhang, Y., Zhang, J., & 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.

Zhou, X., Jiang, X., Ye, Z., Zhang, Y., Lou, K., & Zhan, W. (2010). Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study. *Neuropsychologia*, *48*(6), 1551-1562.

Article 2

Classifier animacy cues distinct gap expectations in head-final relative clause structure building in Mandarin Chinese[1]

Zirui Huang and E. Matthew Husband

University of Oxford

Mailing address:
E. Matthew Husband
St. Hugh's College
St. Margaret's Rd.
Oxford OX2 6LE
United Kingdom

Email: zirui.huang@ling-phil.ox.ac.uk (Huang), matthew.husband@ling-phil.ox.ac.uk (Husband)

Running head: CLASSIFIER ANIMACY & GAP EXPECTATIONS

Word Count: 10819

[1] This study, mainly Experiment 1 and Experiment 2, was conducted as an M.Phil research project and was submitted to University of Oxford as dissertation for the completion of an M.Phil degree in Linguistics, Philology and Phonetics.

Abstract[2]

Relative clause constructions in Mandarin Chinese are prenominal, which makes it particularly challenging to actively anticipate a gap site in the absence of a head-noun filler. Thus, it remains unknown what type of information the parser utilizes to anticipate the structure of an upcoming relative clause in head-final relative clauses and how detailed such structure building is before receiving information from the head noun directly. To address this, we investigated how semantic information in the form of animacy cues provided by different classifiers in Mandarin Chinese guides the structure building of upcoming relative clauses. In a series of studies, we manipulated classifier types (human, non-human, general) to examine whether the parser uses classifier information to predict the gap site in a head-final relative clause. Our results suggest that the semantic properties of classifiers can help the parser to make structural predictions in head-final RC processing before accessing the head noun. In particular, non-human classifiers guide the parser away from the preferred subject-gapped RC structure, facilitating a null subject object-gapped analysis.

Key words: relative clause; classifier; animacy; prediction; filler-gap dependency

[2] Abstracts of this study have been submitted for academic conferences including: the 26th Architectures and Mechanisms for Language Processing (AMLap) and the 34th annual CUNY conference on human sentence processing.

1. Introduction

The processing of relative clauses has been one of the most intensively investigated topics in the field of psycholinguistics, due to their structural complexity which involves establishing a dependency between a modified noun phrase, i.e., the filler, and its grammatical position in an embedded clause, i.e., the gap. Much of this research has been conducted under a framework of active gap filling (Fodor, 1978; Crain and Fodor, 1985; Stowe, 1986; Frazier, 1987), which proposes that the parser actively posits gap sites where they are grammatically permissible when processing filler-gap dependencies. Active gap filling is also often proposed to be a filler-driven process (Frazier and Flores d'Arcais, 1989). The active search for a gap position is initiated after a parser identifies a filler (e.g., a wh-question word), perhaps reflecting an eagerness to resolve the filler-gap dependency to alleviate the burden of maintaining an unresolved filler in working memory (e.g., Chen, Gibson, & Wolf, 2005; Fiebach, Schlesewsky, & Friederici, 2002).

However, a filler need not necessarily occur before its gap, and head-final relative clause constructions are common in languages like Mandarin Chinese. An example of a Chinese head-final RC is given in (1a).

(1) a. A head-final RC construction in Mandarin Chinese:

[Jing cha	zai	yi yuan	kan jian $\_\__i$]$_{RC}$	de	na ge	nv hai$_i$	zheng zai	da	dian hua
police	in	hospital	see	REL	that CL	girl	now	make	phone call
			GAP		FILLER				

The girl who the police saw in the hospital was making a phone call.

b. A main clause construction in Mandarin Chinese:

Jing cha zai yi yuan kan jian na ge nv hai zheng zai da dian hua

police in hospital see that CL girl now make phone call

The police saw in the hospital that the girl was making a phone call.

In (1a), both the head noun '*nv hai*' "girl" and the relativizer '*de*', which indicates the RC boundary, appear after the RC, enclosed by brackets. Importantly, this RC structure is temporarily ambiguous between a main clause, as shown in (1b), and an RC, until encountering the relativizer '*de*'. Comprehenders generally have a preference for a main clause analysis over an RC due to the increased processing complexity for an RC analysis (e.g. Frazier, 1987; Kimball, 1973), and may therefore initially misanalyze the RC as a main clause, especially given robust evidence that parsers are highly incremental (Altmann & Kamide 1999, Crain and Fodor 1985, Frazier and Clifton 1989, Kamide, Altmann, and Haywood 2003, Miyamoto 2002, Traxler and Pickering 1996). Such a head-final RC construction poses a great challenge for the filler-driven active gap filling strategy since there is no wh-phrase to indicate the RC boundaries and no filler NP to initiate the search for its canonical gap.

Relative clauses can, however, also be discontinuous in Mandarin Chinese. In these constructions, the classifier, which is dependent on the head noun of the filler, can be stranded to the left of the RC while the relativizer and head noun remain to the RC's right. Incrementally, this can provide a grammatical cue to the presence of the RC when the classifier and its following word mismatch, as in (2a-b), potentially

permitting comprehenders to recognize the presence of an RC before processing the relativizer. In (2a), the dislocated classifier “wei” can only modify human nouns and its head noun “jizhe” (journalist) is located after the RC. The immediate adjacent word is “jushi” (rock), which is a mismatched noun to the classifier. This mismatch can be a potential cue for the comprehenders that the classifier-noun phrase is divided by an RC. (2b) demonstrates the classifier mismatch in a different type of RC, object-gapped RC. The word next to the classifier “kuai” is a verb “zadao” (hit), creating a category mismatch, and thus can also lead to the disambiguation of the sentence structure.

(2) Mismatched classifier when with an inserted subject-gapped RC:

a. na wei [jushi zadao] de jizhe jingtide huangu sizhou.
That $CL_{(human)}$ rock hit REL journalist cautiously look surroundings
The journalist that the rock hit looked about his surroundings cautiously.

Mismatched classifier when with an inserted object-gapped RC:

b. na kuai [zadao jizhe] de jushi zhangzhe qingtai.
that $CL_{(rock)}$ hit journalist REL rock grow moss
The rock that hit the journalist is covered with moss.

Previous studies in Chinese RC (Hsu 2006, Wu et al, 2009, Chen et al, 2013; Wu et al, 2014) have demonstrated that comprehenders are able to utilize this mismatch cue to anticipate the presence of an RC, suggesting that the processing of head-final RCs can happen even without decisive information from the substantive part of filler (i.e., nouns) and the relativizer. For example, in the self-paced reading experiment

conducted by Wu et al. (2009), faciliatory effects at the adverbs after the head nouns ("jingtide" as in (2a)) and at the main verbs ("huangu" as in (2a)) were found in the sentences with dislocated mismatched classifiers compared with the sentences without the dislocated classifiers. These results suggest that discontinuous classifier constructions allow comprehenders to recognize the presence of a head-final RC early in processing. What is unclear, however, is how much structure comprehenders are able to anticipate with these relatively abstract grammatical cues. Does the parser actively posit different gap sites depending on the grammatical information of a classifier? The goal of this paper is to demonstrate that even without a filler, the (in)animacy feature of classifiers can not only guide the parser's early disambiguation of an ambiguous RC but also generate expectations for particular gaps inside the RC.

The paper is structured as follows. We first overview the active gap filling framework and some of the key studies that investigate the role of classifiers in head-final RC processing. Then we present three experimental studies, one sentence completion task, one eye-tracking study and one self-paced reading study, and talk in detail about our experimental results in the discussion section.

Filler-gap dependency in relative clause and active gap search

RC constructions allow a noun phrase (NP) to be modified by an entire clause. To process sentences containing such a structurally complex form, comprehenders must be able (a) to recognize the boundary of RC, distinguishing the embedded clause from the main clause and (b) to establish a link between the modified NP and its

grammatical position inside the RC. For head-initial RC constructions like in English, a filler NP appears before the embedded clause and sometimes there is also a relativizer explicitly marked as an RC boundary, making the process of recognizing an embedded RC much easier. Then for the establishment of the link between the NP and its canonical site within the RC, as known as the filler-gap dependency, it has been proved to be an active search process. The parser postulates a gap site in advance of sufficient bottom-up input confirming that analysis (Crain & Fodor, 1985; Frazier & Clifton, 1989; Frazier, 1987; McElree & Griffith, 1998; Omaki et al., 2015; Parker, 2017; Pickering & Traxler, 2003; Staub, 2010; Stowe, 1986; Traxler & Pickering, 1996). Early self-paced reading studies (e.g., Crain & Fodor, 1985) demonstrated slower reading times at regions where gaps were expected but absent, suggesting that comprehenders predict gaps and experience processing difficulty when their predictions fail. An eye-tracking while reading study by Traxler and Pickering (1996) manipulated the plausibility between a displaced filler and the verb within the RC, and found longer reading time at the verb when the filler is a semantically implausible object for the verb compared with when the filler is a semantically plausible object. The observed plausibility mismatch effect demonstrated that comprehenders have built a sufficiently detailed structure in which the filler takes on the thematic role of Theme of the verb and therefore confines its gap to the direct object position. That this effect could be observed as early as at the verb indicates that the active expectations for a gap site is rapidly initiated after the parser detects a filler. In this sense, the parsing of filler-gap dependency involves a predictive mechanism as postulating a gap site in

advance requires top-down knowledge.

Investigation focus has also been given to the detailed structure building during active gap search, like, where comprehenders posit the gap sites. Some studies show a general preference for subject gap over object gap (Wanner & Maratsos, 1978; Ford, 1983; Holmes & O'Regan, 1981; King & Just, 1991; Lee, 2004). Lee (2004) observed a filled-gap effect at the subject position in conditions where a subject gap is grammatically licensed compared with conditions where a subject gap is grammatically prohibited. This subject-gap preference might reflect the parser's eagerness to complete the filler-gap dependency due to the working memory burden as subject-gapped RCs involve shorter distances (King & Just, 1991; Gibson, 1998). Moreover, some studies suggest that active gap filling operates with such eagerness that it disregards bottom-up lexical-semantic information about the gap sites (Pickering and Traxler 2003; Staub 2007; Omaki et al., 2015). In Pickering and Traxler's self-paced reading and eye-tracking studies in 2003, they use optionally transitive verbs which are more frequently used as intransitive verbs to test whether the parser still misanalyses the filler as the object of the verb. The results suggest that the parser predicts the gap ahead of the verb, insensitive to bottom-up lexical inputs that indicate the plausibility of a gap site. Furthermore, Omaki et al. (2015) demonstrate how hyper active the gap filling can be by showing the parser attempts to posit a gap at the direct object position for the filler even with strictly intransitive verbs. It suggests that the active search for the gap is triggered before consulting verb subcategorical information. These findings indicate that active gap filling prioritizes dependency formation over

integrating detailed lexical-semantic cues, highlighting the proactive nature of real-time language processing.

However, some semantic factors can indeed modulate the eager active gap filling. For example, the animacy of a filler has been shown to modulate the parser's expectation for RC type (Mak, Vonk & Schriefers, 2002, Traxler, Morris & Seely, 2005, Gennari & MacDonald, 2008). In a sentence completion task in Gennari and MacDonald's study (2008), they found that people tend to complete an RC as an object-gapped RC with given an inanimate filler NP compared with an animate filler NP, which they complete as subject-gapped RC. In their self-paced reading experiment, they find that processing difficulty emerges as early as the relative clause subject when given an animate filler NP. Lowder & Gordon (2014) also find a similar effect in eye-tracking. More recent work by Bovolenta and Husband (2023) shows evidence that subject noun animacy can guide comprehenders to predict different verb phrase structures. For example, an inanimate subject predicts that the subject is derived, hence more object relative clauses will be predicted. Based on these results Gennari & MacDonald (2008, 2009) proposed the Production-Distribution-Comprehension account, arguing that an animate filler NP leads to an expectation of a subject gap while an inanimate filler NP induces a stronger expectation of an object gap. Their theory can be grouped under the label of "expectation-based" theories, which also include the word-order frequency theory (Bever, 1970; MacDonald & Christiansen, 2002), surprisal/expectation framework (Hale, 2001; Levy, 2008), and entropy-reduction accounts (Hale, 2003). The central idea of expectation-based theories is that

comprehenders dynamically adjust the likelihood of the upcoming linguistic inputs and generate up-to-date predictions based on the structure or features demonstrated in the previous inputs. The observation that the animacy of the filler NP can affect gap preference fits the prediction of expectation-based accounts.

Filler-gap dependency in Chinese head-final relative clauses

In Chinese RC constructions, which are prenominal, the relative order between the filler and gap is reversed: the gap appears before the filler. This typological difference in filler-gap position may potentially affect how the parser perceives RC structures. In head-final RC constructions like Chinese, active gap filling, which is assumed to be initiated by the filler in the previous studies on head-initial RCs, is not applicable as the filler is not available for the parser to initiate any gap search. Moreover, a relativizer that explicitly indicates the beginning of an RC is also absent in head-final RCs. Instead, The relativizer marks the right boundary of a prenominal RC, occurring right before the head noun. Consequently, whether and how the parser can make any prediction about an upcoming RC becomes an intriguing question to investigate.

Despite the structural difference between head-final RCs and head-initial RCs, the dependency relations between the gap and the head (filler) in head-final RCs still remain valid as in head-initial RCs. In Chinese head-final RC, there is processing evidence from a classic self-paced reading study by Hsiao & Gibson (2003) showing that object-extracted relatives are processed more easily than subject-extracted relatives in Chinese, the reverse of the English pattern. Their design and analysis tie

the processing difficulty to dependency length and the locus of integration: because Chinese RCs are prenominal, the head noun appears after the clause, so a subject-extracted RC carries a longer filler gap dependency distance and thus heavier integration/storage demands at the relativizer *de* and head-noun region than an object-extracted RC. This localization cost demonstrates the processing trace between the external head and the grammatical position inside an RC. Packard, Ye & Zhou (2011) also report converging EEG results. They found larger P600 effects for subject-extracted RCs than for object-extracted RCs at the relativizer and the head noun. They interpret the P600 effects as the cost of linking the head to the grammatical gap inside an RC. Furthermore, a series of studies conducted by Kwon and colleagues (2008, 2010, 2013) established how comprehenders create and resolve filler gap dependencies in Korean head-final structures. Though Korean has overt case making system while Chinese does not, the head-final structural features are comparable between the two languages. In their eye-tracking experiment (2010), Kwon et al. found a robust subject-extracted RC advantage in Korean and argued that cue-based parsing (case, animacy, structural frequency) guides early gap projection and modulates later integration cost at the head noun—i.e., the parser behaves as if a gap is actively posited and then linked to the upcoming head. In ERP, Kwon et al. (2013) showed asymmetry P600 effects in subject/object RCs and interpreted the effects as structural integration and modulation by case cues, again consistent with the active formation of a syntactic filler-gap dependency. Taken together, in head-final RC construction, early gaps are licensed based on available cues even without head

nouns, and there are processing costs for forming and integrating a syntactic dependency between the gap and the filler demonstrated by timing signatures and asymmetries when processing different types of RCs.

Role of classifier in head-final relative clauses

What linguistic information is available for the parser to license a gap in a prenominal RC when the head is not yet available? One sensible way to probe into this issue is to examine cues that can indirectly indicate RC boundaries and investigate whether they can be utilized by the parser to detect RCs in advance. As mentioned earlier in the introduction, a head-final RC could be temporarily ambiguous between a main clause and an RC construction before the appearance of the relativizer "de". However, the mismatch between linearly adjacent words can potentially indicate the boundary between the main clause and the embedded RC. For instance, when an NP is modified by a classifier and a prenominal RC at the same time, the classifier can surface in front of the RC or after, as shown in (3a) and (3b). The classifier is semantically compatible with its head noun, meaning a classifier can only modify a specific type of nouns. This close relation between a classifier and its head noun has been attested with much empirical evidence (Huettig et al., 2010, Zhang et al., 2012, Zhou et al., 2010). Thus, in cases like (3b) where an RC comes between a classifier and a noun, the mismatch between the classifier ("*ben*" CL_{book}) and the immediately following word ("*peng you*" "friend", a human noun) inside the RC, provide a cue that an RC is inserted in between the classifier and the head noun, and the head noun for

the classifier will come after the RC.

(3) Classifier immediately preceding its head noun:

a. [peng you tui jian] de zhe ben shu

Friend recommend REL this $CL_{(book)}$ book

The book that a friend recommended.

Classifier dislocated from its head noun:

b. Zhe ben [peng you tui jian] de shu

This $CL_{(book)}$ friend recommend REL book

The book that a friend recommended.

How reliable is this classifier mismatch cue to serve as a cue for an upcoming NP structure that includes an RC? We first need to validate the syntactic relation between the classifier and the noun, and then discuss the syntactic possibilities for classifier mismatch. A widely adopted point of view for the classifier-noun relation is that it is strictly local (Cheng & Sybesma 1999; Zhang 2013; Jiang, Jenks & Jin 2022). A classifier is treated as part of the extended nominal spine: Demonstrative > Numeral > Classifier > NP. Under this view, a classifier is a functional head that selects an NP complement and is itself selected by a numeral or demonstrative. A classifier composes with its sister NP inside the nominal domain and cannot cross over intervening clausal material to find another noun. Given this locality constraint, when

the word immediately following a classifier does not satisfy its selectional requirements, i.e., a semantically compatible noun, for example, a verb or a semantically incompatible noun, since the classifier cannot project across clausal boundaries or re-attach to some distant DP, the only syntactically legitimate continuation is that the classifier and its head noun remain within the same DP, with some intervening material inside that DP, i.e., a modifying RC as demonstrated in (3b). Thus, a classifier mismatch necessarily signals that a prenominal RC is inserted between the classifier and its head noun, and that the classifier will ultimately attach to the head noun following the RC.

To further justify the use of the classifier mismatch configuration as an unambiguous predictive cue for the upcoming RC structure in Chinese, other possible continuations following a classifier mismatch that do not lead to an RC have to be examined and eliminated. To start with, classifier floating, which allows classifiers to appear clause-internally and be linearly separated from their host NP, is commonly found in other classifier languages such as Japanese and Korean. However, it is not commonly allowed in Mandarin Chinese. There are some cases in Mandarin Chinese where classifiers appear to be “stranded” and thus seemingly creating a classifier mismatch configuration without leading to an RC construction, as shown in (4a) and (5a). However, these cases are typically nominal-internal configurations like ellipsis, or topicalization in particular discourse, not truly clausal-level classifier floating. (4a) shows the focus fronting of a nominal quantity phrase with NP ellipsis. In replies to the “how many” question, the fronted Numeral–Classifier phrase bears a narrow focus

(answering precisely the quantity) and thus the NP is elided. Mandarin allows for focus-fronting in the left periphery, and a left-edge operator binds an empty internal argument inside the VP, see (4b) for the syntactic structure. (5b) is a demonstration of the topicalization of NP. In this Topic-comment structure, the NP is a left-dislocated topic (aboutness topic), and the NP inside the comment clause is elided and coindexed with the topic. Thus, the split between the noun and its classifier is induced by topicalization rather than by a classic clause-internal floating classifier construction. Such focus-fronting and topicalization are discourse-driven operations that are licensed only in specific information-structural contexts (e.g., contrast or emphasis), and are dispreferred in neutral sentences.

(4) Focus-fronting with NP ellipse:

a. Q: ni chi le ji wan fan?

You eat PERF how many CL rice

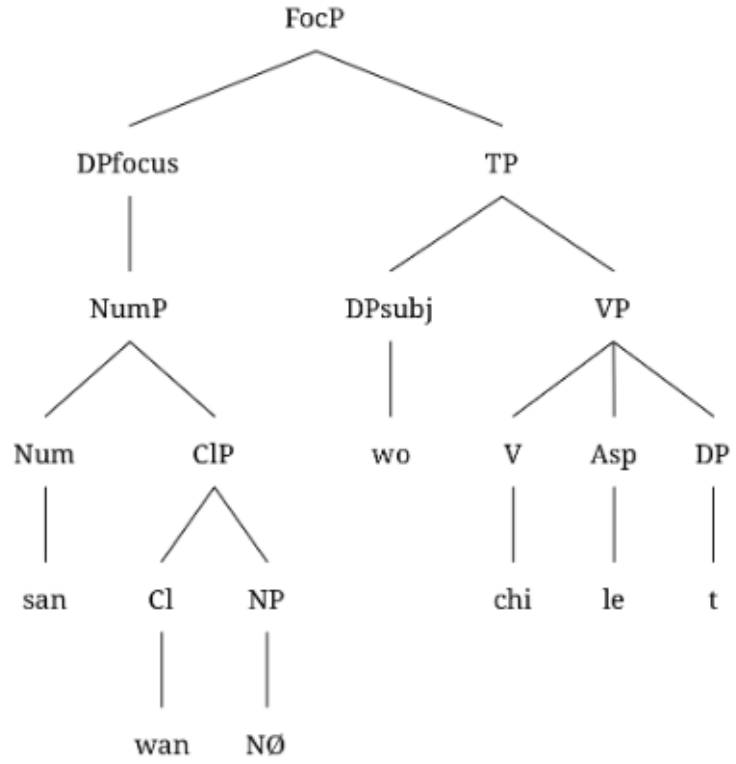
How many bowls of rice did you eat?

A: san wan wo chi. le.

Three CL I eat PERF

Three bowls (of rice) I ate.

b. Syntactic structure of 4a



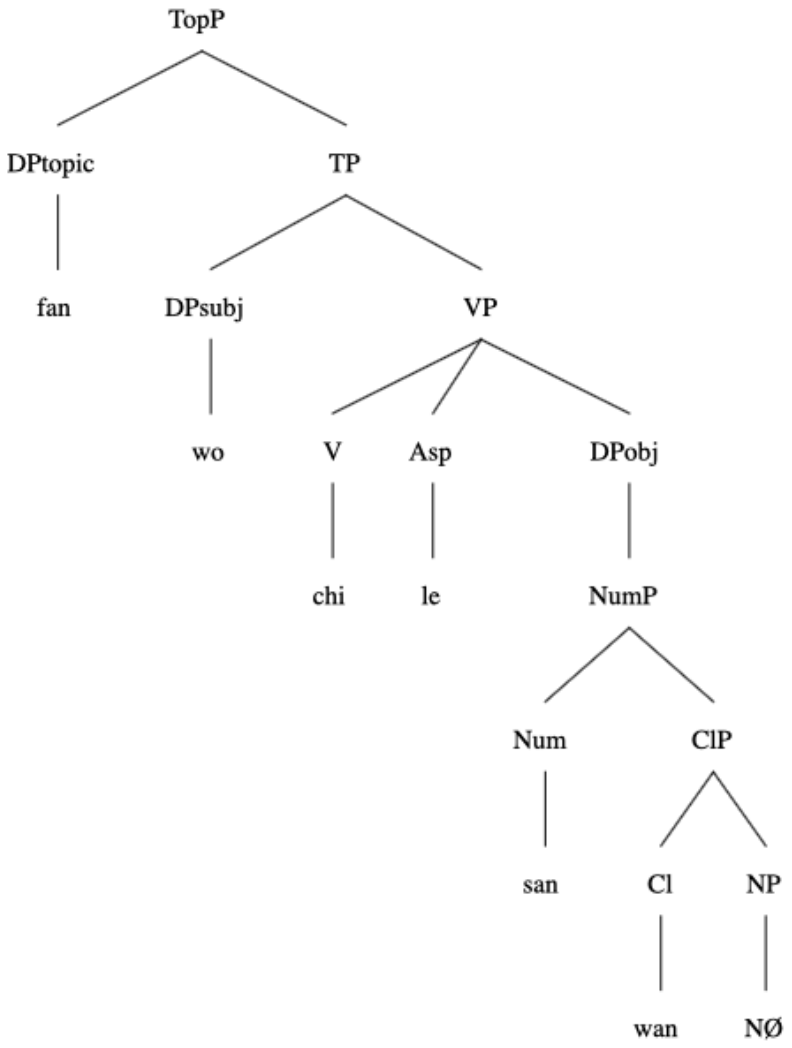
(5) Topicalization of NP:

a. Fan wo chi le san wan

Rice I eat PERF three CL

For rice, I ate three bowls (of it).

b. Syntactic structure of the topicalization of NP



A further logical possibility is that a classifier might modify a null pronoun (pro), and such a construction, in principle, appears without an overt noun and hence creates a mismatch between a classifier and the following word. However, there is no theoretical or empirical need to posit a special “classifier + pro” configuration in Mandarin Chinese to explain “numeral+classifier” strings that surface without an overt noun. “Classifier + pro” would place a referential null pronoun in the complement of the classifier, which is theoretically uneconomical and, empirically, collapses into the well-attested NP

ellipsis account (e.g., Cheng & Sybesma 1999; Zhang 2013; Jiang, Jenks & Jin 2022; Saito, Lin & Murasugi 2008). Besides, as mentioned earlier, a classifier is a functional head in the nominal domain: it selects an NP and itself is selected by a numeral or a demonstrative. Thus, allowing a classifier to modify a DP or assuming a null pro violates standard local selection and category-selection assumptions: the classifier does not combine with DP/pro, but with N/NP, which may be silent under ellipsis.

Besides theoretical accounts, there is already empirical evidence of head-final RC studies supporting that a mismatched classifier can be effectively detected and used as a cue to predict RC structure, both in Japanese (e.g., Phillips & Lau, 2004; Yoshida, 2006; Yoshida, Aoshima & Phillip, 2004) and in Mandarin Chinese (Hsu 2006, Wu et al, 2009, Chen et al, 2013; Wu et al, 2011, 2014, 2018). A self-paced reading study reported in Hsu (2006) found a facilitation effect of classifier mismatch for predicting RC structure when presented with RC-biased discourse contexts, i.e., contexts which introduce two referents and create the need to use modifiers like RC constructions to distinguish the two previously mentioned referents apart. Furthermore, Wu et al. (2009) conducted a series of self-paced reading studies to examine whether classifier mismatch can serve as a reliable cue to facilitate RC prediction in isolated sentences and also compared between subject-gapped RC and object-gapped RC. They found facilitatory effects of mismatched classifiers in both subject-gapped RCs and object-gapped RCs without the support of contexts. Shorter reading times were found at the adverbs after the head nouns (“jingtide” as in (6a)) and at the main verbs (“huangu” as in (6a)) in (6a) and (6c) compared with (6b) and (6d). These results suggest that

discontinuous classifier constructions allow comprehenders to recognize the presence of a head-final RC early in processing. Additionally, they also found that the facilitation effects were numerically larger in subject-gapped RC conditions compared to object-gapped RC conditions, indicating the possibility that pre-RC classifiers might be more effective in subject-gapped RCs. One possible reason is that, a classifier-verb mismatch as in subject-gapped RC is more unambiguous and easier to recognize than the classifier-noun mismatch as in object-gapped RC, thus the classifier mismatch cues are potentially more in subject-gapped RCs. Converging evidence for the classifier mismatch effects also comes from the visual world eye-tracking paradigm by Wu and colleagues in 2014, which shows that comprehenders can use mismatched classifier cues to anticipate a correct RC parse.

(6) Object-gapped RC with classifier:

a. na wei jushi zaidao de jizhe jingtide huangu sizhou.
That $CL_{(human)}$ rock hit REL journalist cautiously look-about surroundings
The journalist that the rock hit looked about his surroundings cautiously.

Object-gapped RC without classifier:

b. jushi zadao de jizhe jingtide huangu sizhou.
rock hit REL journalist cautiously look-about surroundings
The journalist that the rock hit looked about his surroundings cautiously.

Subject-gapped RC with classifier:

c. na kuai zadao jizhe de jushi zhangzhe qingtai.
that $CL_{(rock)}$ hit journalist REL rock grow moss
The rock that hit the journalist is covered with moss.

Subject-gapped RC without classifier:

d. na kuai zadao jizhe de jushi zhangzhe qingtai.
hit journalist REL rock grow moss
The rock that hit the journalist is covered with moss.

(Example sentence from Wu et al., 2009)

The studies mentioned above demonstrate that the parser is able to use mismatched classifiers to facilitate the processing of recognizing an RC domain. However, not much attention has been given to whether a classifier can also play a role in modulating the parser's expectancy of a gap after realizing the presence of an RC. How the filler-gap dependency is completed in head-final RCs remains largely unknown. One underlying assumption of the active gap filling strategy when it was proposed is that, it is a filler-driven process. The expectation of an RC and a gap at the earliest possible position is triggered after the parser perceives a filler and the motivation for such expectation is the working memory burden of maintaining a filler. Now head-final RC constructions and the mismatched classifier cues provide a perfect test ground to investigate if the active filler strategy can be generalized to a broader " active dependency strategy", which could be initiated by either a filler or a gap or any usable linguistic cues. Building upon the findings that a mismatched classifier can trigger an expectation of an RC, we ask further questions about whether an

expectation of detailed gap sites can also be triggered, or the processing of active gap search cannot be initiated in head-final RCs due to the absence of an actual filler. Dislocated pre-RC classifiers containing animate features allow us to test this issue. As we reviewed earlier, manipulating the animacy of filler NPs can modulate the parser's RC gap preference, so we wonder if the animacy feature of a classifier can motivate an active prediction of gap sites.

2. The present study

In this study, we probe further into the role of mismatched classifiers, to investigate whether the animacy information provided by classifiers can guide the parser's preference for gap sites in head-final RCs. The semantic feature that a classifier shares with its head noun can potentially be utilized to predict detailed structure building, i.e., which type of RC to expect. For example, a classifier "ben" indicates that the head noun must be a book-like noun phrase like "*shu*" "book", "*xiang ce*" "photo album" or "*za zhi*" "magazine". However, at the same time verbs also constrain the possible thematic roles that a NP can take on. A transitive action verb like "*hu lue*" "ignore" assigns the thematic role of Agent to its subject. In this case, an NP like "*za zhi*" "magazine", which is modified by the classifier "*ben*", cannot take on the Agent role for the verb and therefore cannot be an eligible subject. So a classifier which suggests that the NP is inanimate can potentially guide the comprehenders to avoid positing a subject gap in an RC. Furthermore, it is possible to have a null subject structure in Mandarin Chinese though it might be dispreferred, as shown in (7a), thus it would still be grammatical for comprehenders to assign the thematic role of Patient to an inanimate NP even the Agent role is not occupied. Such a null subject RC as in (7a) is temporarily ambiguous with a subject-gapped RC reading as shown in (7b), before encountering the relativizer "de" and the head noun. In an object-gapped RC with null subject (ORC-NS): "demonstrative + classifier + verb + REL + head noun", both the subject and object of the RC are null. The subject of the RC is a pro and the object is a gap as the dependent of the filler NP.

(7)

a. Object-gapped relative clause with null subject

Na ge hu lue de xiansuo

That CL ignore REL clue

The clue that (someone) ignored.

b. Subject-gapped relative clause

Na ge hu lue xian suo de jing guan

That CL ignore clue REL police officer

The police officer that ignored the clue.

In this study, we use the mismatch between classifiers and transitive action verbs in ORC-NS constructions as shown in (8a-c). We distinguish three types of classifiers: human classifiers, non-human classifiers, and a general classifier, thus creating three conditions. Human classifiers only modify human NPs like "*xiao hai*" "child", "*ji zhe*" "journalist". The general classifier "*ge*" does not contain specific animacy features about its head nouns and can basically modify any type of nouns including the nouns which are usually modified by a specific set of classifiers. For example, human NPs like "*xiao hai*" "child", "*ji zhe*" "journalist" can also be modified by the general classifier "*ge*", in more informal or colloquial situations. The non-human classifiers cover a very wide range of noun selections, however, in this study, we only select the classifiers and nouns that are strictly inanimate and are unable to serve the Agent role for active verbs.

One thing to note here is that we only focus exclusively on the human nouns, which is the most animate group in the Animacy Hierarchy, and the non-living inanimate entities, which rank in the least animate group. We choose only the two extreme ends on the Animacy Hierarchy to maximize the feature contrast in the linguistic encodings and thematic role assignments. Some intermediate categories in the Animacy Hierarchy like *natural force* (e.g., hurricane), which is semantically inanimate but behaves in ways that are more similar to animate categories as it can initiate movement, change course, and cause destruction, injury, and death (Lowder et al., 2015). The thematic roles for natural forces can therefore be Agent. In this study we avoid such categories which are potentially ambiguous in terms of thematic roles and choose only the human category, which ranks top on the animacy hierarchy and often takes on the Agent role and static inanimate categories, which are generally considered to be unambiguous non-living and cannot take on the thematic role of Agent.

In our design as shown in the example items in (6a-c), if the parser can utilize the semantic information of the classifier to posit a gap in a relative clause, it is more likely for the parser to be led to a garden path when it encounters a human classifier with ORC-NS constructions. With a human classifier, the parser will expect a subject-gapped RC structure because the classifier indicates that the filler NP is a suitable subject in an RC, and the parser will posit the gap at the subject position and expect an object noun phrase to show up. When the parser does not encounter an object but a relativizer, it will disconfirm the prediction of a subject-gap analysis and the parser will have to reanalyze the RC structure. On the other hand, if the classifier is a non-

human classifier, the parser might realize in advance that the head noun is an inanimate object and is not eligible to be the subject of the verb. Then the parser might consider a null subject and posit the gap at the object position. Therefore, when the parser reaches the relativizer and the head noun, no garden path effect is expected. The general classifier “ge” may be seen as a baseline condition since a general classifier cannot provide additional information to help the parser determine sentence structure. However, garden path effects are still expected to show up because a subject-gap structure is, in general, preferred. The expectation is that the reading disruption effects would be much greater in the human classifier condition than in the general classifier condition.

(8)

a. Human classifier condition

Na ming jian dao de hai zi yi jing xing guo lai le

That CL find REL child already awake PERF

The child that (someone) found is already awake.

b. General classifier condition

Na ge jian dao de ying bi yi jing zang xi xi de le

That CL find REL coin already dirty PERF

The coin that (someone) found is dirty.

c. Non-human classifier condition

Na zhang jian dao de yin hang ka yi jing huan gei shi zhu le

That CL find REL card already return owner PERF

The credit card that (someone) found has already been returned to its owner

3. Experiments

Experiment 1: Sentence completion

As discussed earlier, there are possible configurations in which classifier mismatch does not necessarily cue an RC construction. Although such configurations either arise only in specific discourse contexts or are theoretically ruled out, it is still important to verify how reliable the classifier mismatch cue is in predicting an RC structure. Thus, a sentence-completion task was conducted to examine (a) whether comprehenders use a mismatch between a classifier and the following verb as a cue to produce a relative clause, and (b) whether a non-human classifier biases comprehenders toward producing more object-gap relative clauses with a null subject.

Participants

439 native Mandarin speakers were recruited on Wenjuanxing, a Chinese online survey platform. They were all given informed consent and received shopping vouchers as compensation for their time. All methods used were approved by the Social Sciences and Humanities Interdivisional Research Ethics Committee and the University of Oxford (Ref. No. R66947).

Material

An example set of items used in the sentence completion task is given in (9a-c). Transitive verbs and experience verbs which allow a null subject are selected. Each verb was combined with a human classifier “ming” or “wei”, the general classifier “ge”, and a non-human classifier such as “ben”, “tiao” or “zhang”, forming 32 experimental items with three conditions. Experimental items long with 192 filler items were

counterbalanced and evenly distributed to 12 lists.

(9) Example set of sentence fragments:

a. Human classifier condition

那 名 喜 欢 _____

Na ming xi huan

That CL_{human} like _____

b. General classifier condition

那 个 喜 欢 _____

Na ge xi huan

That $CL_{general}$ like _____

c. Non-human classifier condition

那 本 喜 欢 _____

Na ben xi huan

That $CL_{non\text{-}human}$ like _____

Procedure

Participants were first presented with a brief introduction, stating that they were expected to complete the sentence with the appropriate content in their mind and there was no time pressure on finishing the list, followed by examples of how to complete sentence fragments. Each participant was randomly assigned with one of the twelve lists, manipulated by the built-in randomization program of the survey platform. The estimated time for completing a list was 2-5 minutes.

Results

3512 responses were collected and analyzed. In 69.3% (2434 out of 3512) participants produced relative clause structures when “classifier + verb” is given. 21.9% (768 out of 3512) were invalid answers including ungrammatical sentences, claims that they were not sure how to complete the sentence and nonsensical responses. 8.8% were grammatical sentences other than relative clauses. This result shows that the “classifier + verb” prompt effectively elicited relative clause constructions (X^2 = 2136.2, df = 2, p < .001). Among the responses that produced relative clauses, 1452 of them (59.7%) were subject-gapped relative clauses, and the remaining 982 responses (40.3%) were object-gapped relative clauses. In human classifier condition, 92.2% (720 out of 781) were subject-gapped relative clause, which is predominately higher than object-gapped relative clauses (X^2 = 556.06, df = 1, p < .001). in general classifier condition, 71.4 % were subject-gapped relative clauses which is also significantly higher than the object-gapped relative clauses (X^2 = 158.7, df = 1, p < .001). However, in non-human relative clause conditions, only 14.1% were subject-gapped relative clauses which is significantly lower than object-gapped relative clauses (X^2 = 403.89, df = 1, p < .001). Most participants (85.9%) gave object-gapped relative clauses when seeing a non-human classifier. A summary of counts and percentages of types of responses is presented in Table 1.

Structure / Condition	Relative clause		Other structures	
	Subject-gapped	Object-gapped	Others	Grammatical
Human CL	781 (67.0%)		384 (33.0%)	
	720 (92.2%)	61 (7.8%)	276 (71.9%)	108 (28.1%)
General CL	871 (66.1%)		447 (33.9%)	
	622 (71.4%)	249 (28.6%)	357 (79.9%)	90(20.1%)
Non-human CL	782 (76.0%)		247 (24.0%)	
	110 (14.1%)	672 (85.9%)	135 (56.6%)	112 (49.4%)
Total	2434 (69.3%)		1078 (30.7%)	
	1452 (59.7%)	982 (40.3%)	768 (71.2%)	310 (28.8%)

Table 1. Counts and percentages of different types of responses

Discussion

From an overall view, a majority of native Mandarin speakers (69.3%) produce a relative clause construction for classifier + verb constructions, suggesting that the mismatch between a dislocated classifier and its following verb was an effective cue to indicate that a relative clause was expected to occur in between the classifier and the head noun. As for the 8.8% grammatical responses which are non-relative-clause structures, they can be categorized into two types. In the first type, participants treated the classifier as a noun and they were not expecting another noun as a head. Therefore, no relative clause structure is produced. These classifiers that can function as nouns and elicit non-relative-clause structures were excluded from the experimental material

for the following eye-tracking and self-paced reading studies. Another type of non-relative-clause structure occurred when participants treated the verb as the head noun. Some words in Chinese can be used in different word categories. Due to the lack of an inflection system, there is no morphological difference for a word when it is in different syntactic categories and the interpretation of the syntactic category of a word is often largely context-dependent. Therefore, some participants considered the word following a classifier as the head noun instead of a verb, and no relative clause was elicited. The verbs that can also be used as nouns were also excluded from the experimental material for the following eye-tracking and self-paced reading studies.

The proportions of subject-gapped relative clause (71.4%) and object-gapped relative clause (28.9%) in general classifier condition further confirm the hypothesis that subject-gapped relative clause structure is preferred over object-gapped relative clause with null subject. In human classifier condition the preference for subject-gapped relative clauses grew (92.2%) while in non-human classifier condition, the proportion for subject-gapped relative clauses dropped dramatically (14.1%), suggesting that people can utilize and be guided by the semantic features of various classifiers to predict the semantic features of the head nouns and thus predict different gap site in relative clauses.

Experiment 2: Eye-tracking study

The aim for the eye-tracking while reading experiment was to examine whether the parser utilize the semantic information provided by different classifiers (human,

non-human and general) in real-time processing to guide structure building of upcoming relative clauses.

Participants

42 native Mandarin Chinese speakers were recruited in Oxford (age: 19 to 56; 29 females and 13 males). All of them reported normal or corrected-to-normal vision. The average time for the whole process of this experiment was around 45 minutes and participants received £8 cash as compensation for their time.

Material

Materials consisted of 12 experimental items and 24 filler items. In the experimental items, we used object-gapped relative clauses with null subject and paired them with three different types of classifiers, creating three conditions: human classifier condition, general classifier condition and non-human classifier condition. An example set is given in (10a-c). These experimental items were selected based on the results of the sentence completion task.

(10) Example set of stimuli：

a. Human classifier condition

那 名 捡到 的 孩子 已经 醒过来 了。

Na ming jian dao de hai zi yi jing xing guo lai le

That CL find REL child already awake PERF

The child that (someone) found is already awake.

b. General classifier condition

那 个 捡到 的 硬币 已经 脏兮兮的 了。

Na ge jian dao de ying bi yi jing zang xi xi de le

That CL find REL coin already dirty PERF

The coin that (someone) found is dirty.

c. Non-human classifier condition

那 张 捡到 的 银行卡 已经 还给 失主 了。

Na zhang jian dao de yin hang ka yi jing huan gei shi zhu le

That CL find REL credit card already return owner PERF

The credit card that (someone) found has already been returned to its owner.

Procedure

Eye movements were recorded with an EyeLink 1000 system (SR Research) at a sampling rate of 1,000 Hz. Calibration and validation were performed at the beginning of each block and throughout the experiment as needed. The participants were asked to read the sentence on the screen and to fixate at a black dot at the top-right corner area of the screen when they finished reading. Then a true/false question related to the previous sentence was displayed and the participants should answer the question

according to their understanding of the previous sentence by pressing a button. One-third of the trials were followed by comprehensive questions and the participants were told to read the sentences carefully and make sure to answer the true/false questions correctly. An introduction of the procedure was given first and then they would enter a practice session with four filler trials. After this practice session, they would go through three experimental blocks in which the experimental trials and the filler sentences were presented in a pseudorandom order ensuring that no more than two consecutive trials are from the same experimental condition.

Data analysis

First fixation, first pass reading time, regression path duration, total dwell time, regression in and regression out were the main measures reported in the online sentence processing and were defined as follows. First fixation duration and first pass reading time are viewed as an early measure of initial sentence processing (Clifton, Staub and Rayner, 2007). Regression path duration and total dwell time are considered to indicate the reading difficulties at more intermediate stages of sentence processing. Fixation points shorter than 80ms were automatically incorporated into larger ones since not much information can be extracted from such short fixations (Rayner and Pollatsek, 1989). All these measures were processed and calculated by DataViewer.

Six interest areas (IA) were defined: the determiner (Det), the classifier (CL), the verb(V), the relativizer (REL), the head noun(N), and the spill-over. The REL and N

regions were the critical areas where reading disruptions were expected to occur in the human classifier condition and reading facilitations in the non-human classifier condition. The general classifier was taken as the baseline condition to compare the effects of the human classifier and non-human classifier. Statistical analyses were carried out over the eye-tracking measures mentioned above in these regions using linear mixed-effect models (Baayen et al., 2008). The mixed effects models included random intercepts for participants and for items.

Results

Table 2 presents the participant means on each measure for each region in milliseconds, and Table 3 presents a summary of model estimates and t-values for six eye movement measures. In the Det region, a longer total reading time was found in the human classifier condition compared with the baseline general classifier condition (Est. =14.515ms, t=2.692, p=0.007). No effects were found in any other measure in either the human classifier condition or the non-human classifier condition. In CL region, the first fixation (Est.=20.310ms, t=3.619, p<0.001), first pass reading (Est.=20.374ms, t=3.155, p=0.002), regression path duration (Est.=17.930ms, t=2.131, p=0.040) and total reading time (Est.=54.073ms, t=4.811, p<0.001) all showed main effects of human classifier. Reading time was longer when with a human classifier than when with the general classifier. No effect showed up in the non-human classifier condition in any measure. In V region, longer time in the measures of regression path duration (Est.=33.814ms, t=2.042, p=0.049) and total reading time (Est.=117.610ms,

t=3.809, p<0.001) were observed in human classifier condition. The verb region also received more regression-ins from the later region in the human classifier condition (Est.= 0.053, t=3.038, p=0.004) and fewer regression-ins in the non-human classifier condition (Est.=-0.036, t=-2.155, p=0.038).

In the REL region, which is the critical region, regression path duration (Est.=66.300ms, t=3.499, p=0.001) and total reading time (Est.=46.590ms, t=3.969, p<0.001) revealed longer fixation time in human classifier condition. More importantly, evidence for significant reading facilitation in non-human classifier showed up in the measures of first fixation (Est. =-12.241ms, t=-2.399, p=0.021), first pass reading (Est.=-14.172ms, t=-2.545, p=0.015), regression path duration (Est.=-39.380ms, t=-2.077, p=0.045) and total reading time (Est.=-48.620ms, t=-4.139, p<0.001) with shorter fixation time compared with baseline condition. Figure 1a presents the comparison of different conditions in first-pass reading and regression path duration at the REL region. In the N region, which was also a critical region, significant differences emerged in human classifier condition in the measures of first pass reading (Est.=14.337ms, t=2.326, p=0.026), regression path duration (Est.=86.640ms, t=4.310, p<0.001) and total reading time (Est.=58.670ms, t=2.390, p=0.022) with longer reading time. In the non-human classifier condition, a significant effect showed up in regression path duration (Est.=-58.270ms, t=-2.842, p=0.007) and total reading time (Est. =-81.350ms, t=-3.314, p=0.002) with shorter fixation and also a marginal effect in first-pass reading (Est.=-11.868ms, t=-1.898, p=0.065) compared with baseline condition. Figure 1b shows the comparison of different conditions in first-pass reading and

regression path duration at the head noun region.

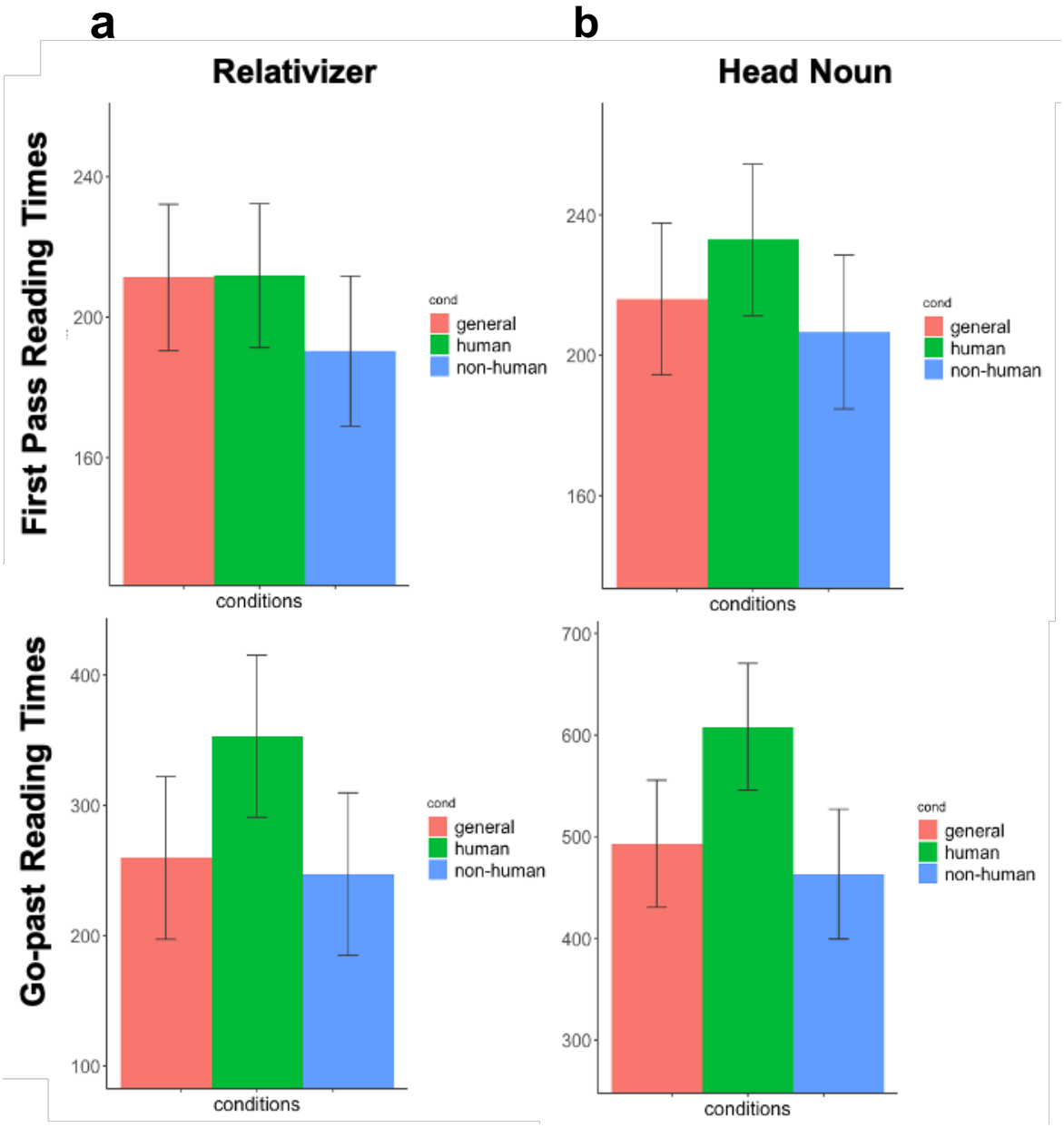


Figure 1. First pass reading time and go past reading at the relativizer and the head noun region.

Measure	that	CL	V	REL	N	spill
First fixation						
General CL	229.87	231.03	233.87	208.77	197.30	218.78
Human CL	267.43	277.52	234.46	207.67	209.03	225.35
Non-human CL	259.80	250.93	228.42	186.76	198.15	215.86
First pass						
General CL	262.08	241.96	300.92	212.34	215.26	248.59
Human CL	280.59	297.93	324.59	215.81	233.85	257.02
Non-human CL	283.88	272.88	299.94	189.26	206.35	239.65
Regression path						
General CL	262.08	244.84	366.85	237.29	282.27	294.73
Human CL	280.59	301.97	406.66	261.28	319.55	317.54
Non-human CL	291.56	281.68	381.88	220.89	254.15	269.76
Total reading time						
General CL	52.47	149.95	555.39	185.72	377.93	358.49
Human CL	78.05	260.96	727.85	230.14	414.21	415.56
Non-human CL	59.77	209.53	544.25	134.08	271.35	326.75
Regression in						
General CL	0.87	0.92	0.80	0.52	0.37	0.25
Human CL	0.93	0.92	0.87	0.47	0.36	0.27
Non-human CL	0.92	0.93	0.78	0.48	0.35	0.21
Regression out						
General CL	-	0.01	0.11	0.14	0.19	0.14
Human CL	-	0.01	0.11	0.16	0.17	0.14
Non-human CL	-	0.01	0.13	0.13	0.14	0.15

Table 2. Mean reading time in each measure in milliseconds

measure	that		CL		V		REL		N		spill	
	Est.	t	Est.	t	Est.	t	Est.	t	Est.	t	Est.	t
First fixation												
(intercept)	204.70	10.95	248.23	29.69	232.23	25.08	200.16	24.35	202.27	27.55	5.29	150.28
Human CL	46.91	**2.83****	20.31	**3.62*****	2.66	0.58	4.15	0.85	6.84	1.51	0.04	1.02
Non-human CL	46.89	**2.66****	0.01	0.00	-3.86	-0.84	-12.24	**-2.39***	-3.21	-0.70	-0.00	-0.01
First pass												
(intercept)	235.35	10.18	263.83	23.51	305.77	20.71	204.49	22.294	218.51	23.80	5.38	127.94
Human CL	27.38	1.28	20.37	3.15**	16.47	1.88	7.35	1.39	14.34	**2.32***	0.05	0.94
Non-human CL	38.46	1.69	5.54	0.85	-7.99	-0.91	-14.17	**-2.545***	-11.86	-1.898	-0.01	-0.19
Regression path												
(intercept)	5.33	72.44	297.03	20.42	503.02	16.24	286.61	11.22	521.42	20.95	5.88	88.45
Human CL	0.13	**2.19***	17.93	**2.13***	33.81	2.04*	66.30	**3.50****	86.64	**4.31*****	0.08	1.04
Non-human CL	0.15	**2.41***	14.20	1.68	6.29	0.38	-39.38	**-2.07***	-58.27	**-2.84****	-0.01	-0.07
Total reading												
(intercept)	55.16	3.97	208.65	9.47	605.96	13.98	182.94	10.385	353.34	13.80	357.76	7.93
Human CL	25.42	**2.72****	54.07	**4.81*****	117.61	**3.81*****	46.59	**3.969*****	58.67	**2.39***	61.82	1.06
Non-human CL	7.30	0.78	3.00	0.26	-62.03	-2.00	-48.62	**-4.139*****	-81.35	**-3.31****	-29.59	-0.51
Regression in												
(intercept)	0.86	22.51	0.00	42.57	0.80	26.85	0.51	13.75	0.37	9.33	0.24	7.07
Human CL	0.06	1.40	0.00	0.30	0.06	**2.25***	-0.04	-1.09	-0.01	-0.12	0.02	0.46
Non-human CL	0.05	1.05	0.00	0.84	-0.02	-0.73	-0.03	-0.92	-0.01	-0.49	-0.03	-0.95
Regression out												
(intercept)	-	-	0.01	2.03	0.12	7.62	0.14	7.94	0.17	9.87	0.14	5.93
Human CL	-	-	0.00	0.03	-0.01	-0.74	0.01	0.66	0.01	0.32	0.01	0.26
Non-human CL	-	-	0.00	0.47	0.01	1.16	-0.01	-0.66	-0.02	-1.72	0.01	0.42

* p<0.05; ** p<0.01; *** p<0.001

Table 3. Summary of model estimates and t-values for six eye movement measures.

Discussion

The main findings of this eye-tracking study were first, the reading facilitation at the critical region, the relativizer and head noun, in non-human classifier condition compared with the other two conditions; and second, the reading disruption at the relativizer and the head noun in human classifier region compared with the general classifier region. These effects show that when with a human classifier or the general classifier, the parser initially anticipated a subject-gapped relative clause and expected a noun phrase to show up as the object of the relative clause instead of a relativizer. At the relativizer region, we saw an immediate (in early measures like first fixation and first pass reading) and sustain (in later measures like regression path duration and total dwell time) facilitation in the non-human classifier condition, suggesting that a non-human classifier can effectively guide the parser away from the preferred subject-gapped relative clause structure. These reading facilitation effects continued on the head noun region in all reading time measures. While no significant difference showed up between the human classifier and general classifier in early measures, suggesting that in both the human classifier condition and general classifier condition, the parser was led to a subject-gapped garden path, suggesting a general preference for a subject-gap in “CL+verb” combination. Then reading disruption showed up in the human classifier condition in regression path duration and total reading time, suggesting that it was more difficult for the parser to recover from a garden path led by a human classifier than a general classifier. These effects continued to emerge at the head noun region.

These results in general confirm the predictions of this study that the parser was sensitive to the semantic information of different types of classifiers and the cues provided by the classifier were used by the parser to build relative clause structure in advance of encountering a relativizer and head noun. To be more specific, a non-human classifier can guide the parser away from a subject-gapped relative clause. Both general and human classifiers lead the parser to predict a subject gap and cause the initial analysis to end up in a garden path, with a human classifier creating more difficulty in garden path recovery. However, one possible confound is that the head nouns across the three experimental conditions were not controlled as the same nouns. We matched the visual complexity and the word frequency of the head nouns for different conditions to minimalize the confound effect caused by different head nouns, but the animacy of the head nouns and the parafoveal processing might potentially play a role in affecting the effect we found on the relativizer.

Experiment 3: self-paced reading

The aim of this experiment is to replicate the findings in the eye-tracking study with better-controlled material. Since it is impossible to keep the head nouns for human classifiers and non-human classifiers the same, we separately compared human classifiers and non-human classifiers with the general classifier in a two-part experiment.

Experiment 3a: non-human vs. general classifier

Participants

43 native Mandarin speakers were recruited online for the first part of this experiment. They were all given informed consent. The average time for this experiment was around 30 minutes and participants received Chinese currency equivalent to £5 as compensation for their time.

Material

The material consisted of 30 experimental items and 60 filler items. We used the same sentence construction (object-gapped relative clause with null subject) as in experiment 2, and the classifiers were either a non-human classifier or the general classifier, with the head nouns kept the same. An example set of experimental items is given in (11a-b)

(11) Non-human vs. general classifier example items:

a. Non-human classifier condition:

那 **条** 忽略 的 线索 是 破 案 的 关键。

Na tiao hu lue de xian suo shi po an de guan jian

That **CL** ignore REL clue is solve case POSS. key

The clue that (someone) ignored is the key to solve the case.

b. General classifier condition:

那 个 忽略 的 线索 是 破 案 的 关键。

Na ge hu lue de xian suo shi po an de guan jian

That **CL** ignore REL clue is solve case POSS. key

The clue that (someone) ignored is the key to solve the case.

Procedure

The self-paced reading task was built up on the online experiment platform IbexFarm. The experiment began with an instruction and a practice session with 4 filler trials. The experimental trials and filler trials were combined in a pseudo-random order to make sure that no two experimental trials appeared next to each other. Only one word at a time was presented on the screen. Participants pressed the spacebar to see each word in the sentence and the time between each keyboard pressing was recorded.

Data analysis and Results

Reading time for each word-by-word region was examined. The critical region was region 4 (the relativizer) and the following region 5 (the head noun) in which continuous effects might be observed. The critical regions suggested a reading time difference in whether a gap was anticipated or not. Statistical analyses were carried out on reading time across all regions using linear mixed-effect models (Baayen et al., 2008).

Longer reading times at the verb region were found in non-human classifier condition (Est=30.37 ms, t=2.892, $p<0.01$). But in the critical regions, shorter reading time showed up in non-human classifier condition in both the relatizier region (Est=-36.93 ms, t=-3.916, $p<0.001$) and head noun region (Est=-47.27 ms, t=-4.941, $p<0.001$). No significant effects in the spill-over region and the rest regions.

Experiment 3b: human vs. general classifier

Participants

40 native Mandarin speakers were recruited online for the second part of this experiment. All procedures were the same as described in Experiment 3a.

Material

All procedures were the same as described in Experiment 3a. The classifiers were either a human classifier or the general classifier, with the head nouns kept the same. An example set of experimental items is given in (12a-b)

(12) Human vs. general classifier example items:

a. Non-human classifier condition:

那 **名** 忽略 的 证人 是 破 案 的 关键。

Na **ming** hu lue de zheng ren shi po an de guan jian

That **CL** ignore REL passerby is solve case POSS. key

The passerby that (someone) ignored is the key to solve the case.

b. General classifier condition:

那 **个** 忽略 的 证人 是 破 案 的 关键。

Na **ge** hu lue de zheng ren shi po an de guan jian

That **CL** ignore REL passerby is solve case POSS. key

The passerby that (someone) ignored is the key to solve the case.

Procedure

All procedures were the same as described in Experiment 3a.

Results

Longer reading time at the verb was found in the human classifier condition (Est=35.08 ms, t=2.898, p<0.01) Greater reading time differences continue at the critical region relativizer (Est=24.71 ms, t=2.413, p<0.05) and head noun (Est=37.16 ms, t=2.75, p<0.01). No significant effects in the spill-over region and the rest regions.

Discussion

We extended the results we found in the eye-tracking study with an improved split design that keeps the head nouns the same across different conditions. Greater reading time at critical regions (relativizer and head nouns) in human classifier condition compared with general classifier suggests a gap is not posited during processing thus an actual object is expected. Reading facilitations in the critical regions in non-human classifier condition indicates that the parser is guided away from the garden-path of a subject-gapped structure and a gap at the object position is expected or easier to adapt to. Reading time differences at the verb were unexpected. Unlike being distributed in the human classifier condition and being facilitated in the non-human classifier condition, both human and non-human classifiers were showing longer reading time at the verb region. One possible reason is that, compared with a general classifier which does not contain particular semantic information, human classifiers and non-human classifiers might create a larger mismatch effect when encountering the verb. The fact that the reading time dropped at the relativizer region even with longer reading at the verb in the non-human classifier region and reading time differences increased at relativizer and head noun regions in the human classifier

region suggested that the effects at the critical region were not merely a spill-over of the mismatch effect but the results of having different structural expectations.

4. General Discussion

In the present study, an offline sentence completion task, an eye-tracking while reading experiment, and a self-paced reading experiment were conducted to investigate the role of animacy features of classifiers in predicting gap sites in Chinese head-final relative clauses. The sentence completion task revealed that the animacy of classifiers influenced the type of RCs produced. Inanimate classifiers led to the production of object-gapped RCs with null subject though it is generally dispreferred. Animate (human) and general classifiers both elicited predominantly more subject-gapped RCs, aligning with the general subject-gap preference when the classifier is in the pre-RC position. These findings suggest that classifier animacy cues can successfully elicit the production of different RCs in off-line tasks. The real-time experiment results, including an eye-tracking experiment and a self-paced reading experiment, are consistent with the off-line sentence completion tasks, confirming that classifier animacy influences gap site predictions during real-time language processing. Animate (human) classifiers reinforced subject-gap predictions, causing disruptions when processing the relativizer which followed immediately after a transitive verb. While inanimate classifiers guided the parser away from the subject-gap prediction and facilitated the reading of object-gapped RC structure, thus relieving the processing cost at the relativizer. The comparisons were also made with a general classifier which lacks specific animacy features and can be used to modify any type of nouns. Both disruption effects with animate classifiers and facilitation effects with inanimate classifiers were found on the relativizer and the head noun in

the self-paced reading experiment with more controlled sentence materials, solidifying that animacy features drive predictive structure-building for gap sites in head-final RC processing.

This study extended the active gap filling framework to incorporate head-final RC constructions. Previous works on active gap filling (e.g, Crain & Fodor, 1985; Frazier & Clifton, 1989; Stowe, 1986; Traxler & Pickering, 1996; Omaki et al., 2015) are primarily based on the observation in head-initial constructions, where a filler NP provides decisive information for initiating the parser's gap search. However, the availability of filler NP for initializing active gap search should not be taken for granted. In Chinese head-final RC constructions, only a partial filler, in this study, a fronted classifier, can serve as a cue for the upcoming RC and the plausible thematic role for the gap. Given the limited information that classifiers can provide compared with full filler NPs, it gives rise to much more uncertainty in predicting upcoming linguistic materials. Despite this, our results demonstrate that the parser is still willing to commit to a rather detailed RC structure interpretation based on the animacy of classifiers, even when this risky commitment may lead to reanalysis.

A novel contribution of this study is the isolation of animacy feature cues from lexical cues, suggesting that active gap filling can be triggered by features, rather than being entirely filler-driven. The animacy effects in modulating the RC type expectations have been verified in previous studies in head-initial RC constructions (e.g., Mak et al., 2006; Traxler, Williams, Blozis & Morris, 2005). But in these studies, the animacy features are entailed in specific lexical items and they manipulate the animacy feature

by manipulating the filler NPs. Chinese classifiers, by contrast, are less constraining and do not pre-activate specific nouns. Instead, they encode semantic properties (in this study, the animacy feature) of the head nouns without fully specifying them. The results that the parser is sensitive to the classifier animacy feature without access to a specific filler NP provide clearer evidence that a less-constraining semantic feature that is disassociated with specific nouns can also trigger active gap search.

These results challenge theories suggesting that RC structure prediction is solely guided by working memory constraints or general gap frequency preferences, which favor shorter dependency lengths (e.g., subject gaps > object gaps > oblique gaps; Gibson, 1998, 2000). Instead, the parser uses cues like animacy to guide predictive structure building and expecting gaps that are consistent with the classifier animacy. It aligns with an expectation-based account (Levy, 2008; Hale, 2001) in which the parser uses cues like animacy incrementally to make predictions for upcoming structures. The structural differences in these studies are related to the thematic interpretation of the expected gap. Inanimate classifiers in particular are more likely to predict object gaps compared to subject gaps because objects are typically interpreted as Themes, which can be inanimate, while subjects are typically interpreted as Agents, which normally require an animate referent. Signaling an inanimate referent, even without the head noun occurring, is enough to guide the parser away from a subject gap which would receive an Agent interpretation. Animate classifiers, however, are more likely to be related to subject gaps because subjects are typically interpreted to be Agents.

One more thing to note is that the findings in this study cannot speak for the

general preference for subject RC in Mandarin Chinese. A long-standing observation in the field of language processing is the subject RC preference over object RC (Frazier, 1987; Kimball, 1973; Gibson, 1998). It has been established in many languages. However, the RC preference pattern in Mandarin Chinese has been less clear with mixed results from different studies. Lin and Bever (2006) suggest subject RC preference in Chinese, consistent with the findings in other languages because subject RCs have a shorter distance between the gap and the filler. In contrast, other studies report a processing advantage for object RC in Chinese. Hsiao and Gibson (2003) found that object RCs are processed faster especially when the filler is animate. Several studies suggest that RC preferences in Chinese depend on contextual cues or experimental factors. Vasishth et al. (2013) argue that cross-linguistic processing strategies and task design can significantly affect the observed preferences. More relevant to this study, a corpus-based examination conducted by Wu and colleagues (2011) suggested that, when classifiers appear in pre-RC positions, subject RCs appear more frequently. However, the situation is reversed when classifiers occur in post-RC positions and the distribution for object-gapped RC increases. Given the fact that the focus of this study is to use the pre-RC classifiers to elicit structural predictions, our experimental materials are all with pre-RC classifiers and thus naturally biased for subject-gapped relative clauses. It is also reflected in the production results in the general classifier condition that more subject-gapped RCs are produced than object-gapped RCs. Thus our results cannot indicate any general processing preference for RCs.

5. Conclusion

This study investigated the role of classifiers for structure building in head-final RCs in Mandarin Chinese. We have argued that the internal structure building of RCs could happen before approaching the relativizer and the head nouns when a dislocated classifier provides semantic information about the head nouns. We used an object-gapped RC construction with a null subject and manipulated the types of classifiers to examine the effects of classifiers on RC structure building. The results showed that the parser was sensitive to different types of classifiers. Non-human classifiers can effectively guide the parser away from a subject-gapped RC structure while human classifiers led the parser to initially adopt a subject-gapped structure. These results suggested that the parser was strongly predictive even with less constraining semantic feature cues.

References:

Aoshima, S., Yoshida, M., & Phillips, C. (2009). Incremental processing of coreference and binding in Japanese. Syntax, 12(2), 93-134.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language, 59(4), 390-412.

Bever, T. G. (1970). The cognitive basis for linguistic structures. Cognition and the development of language.

Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. Journal of Memory and Language, 52(1), 144-169.

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. Empirical approaches to linguistic theory: Studies of meaning and structure, 43-62.

Cheng, L. L. S., & Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. Linguistic inquiry, 30(4), 509-542.

Clifton Jr, C., & Frazier, L. (1989). Comprehending sentences with long-distance dependencies. In Linguistic structure in language processing (pp. 273-317). Dordrecht: Springer Netherlands.

Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. Eye movements, 341-371.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers. Natural language parsing: Psychological, computational, and theoretical perspectives, 94-128.

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: The processing of German WH-questions. Journal of Memory and Language, 47(2), 250-272.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. Journal of verbal learning and verbal behavior, 22(2), 203-218.

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. Linguistic Inquiry, 9(3), 427-473.

Frazier, L. (1987). Syntactic processing: evidence from Dutch. Natural Language & Linguistic Theory, 5(4), 519-559.

Frazier, L., & d'Arcais, G. B. F. (1989). Filler driven parsing: A study of gap filling in Dutch. Journal of memory and language, 28(3), 331-344.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. Journal of memory and language, 58(2), 161-187.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. Cognition, 111(1), 1-23.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1-76.

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

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. Cognitive science, 29(2), 261-290.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. Journal of Memory and Language, 46(2), 267-295.

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. (2003). The information conveyed by words in sentences. Journal of psycholinguistic research, 32, 101-123.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. Journal of Verbal Learning and Verbal Behavior, 20(4), 417-430.

Huang, Z., (2020). Classifier as a cue for structure building in head-final relative clause in Mandarin Chinese (M.Phil thesis). University of Oxford

Huang, Z., & Husband, E., M. (2021). Classifier as a cue for structure building in head-final relative clause in Mandarin Chinese. In 34th Annual CUNY Conference on Human Sentence Processing.

Huettig, F., Chen, J., Bowerman, M., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. Journal of Cognition and

Culture, 10(1-2), 39-58.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. Cognition, 90(1), 3-27.

Hsu, C.-C. N. (2006). Issues in head-final relative clauses in Chinese -- Derivation, processing and acquisition. Unpublished Ph.D. dissertation, University of Delaware.

Jiang, L. J., Jenks, P., & Jin, J. (2022). The syntax of classifiers in Mandarin Chinese. The Cambridge handbook of Chinese linguistics, (24), 515-549.

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

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

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. Cognition, 2(1), 15-47.

Komatsu, L. K., & Nakajima, Y. (2004). Understanding the parser's choices in visually situated sentence comprehension. Psychonomic Bulletin & Review, 11(5), 890-896.

Kwon, N. (2008). Processing of syntactic and anaphoric gap-filler dependencies in Korean: Evidence from self-paced reading time, ERP and eye-tracking experiments. University of California, San Diego.

Kwon, N., Gordon, P. C., Lee, Y., Kluender, R., & Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in Korean. Language, 86(3), 546-582.

Kwon, N., Kluender, R., Kutas, M., & Polinsky, M. (2013). Subject/object processing asymmetries in Korean relative clauses: Evidence from ERP data. Language, 89(3), 537-585.

Kwon, N., & Sturt, P. (2014). How does structural parallelism influence the online processing of center-embedded sentences? Evidence from ERPs. Journal of Memory and Language, 74, 16-37.

Lau, E. F., & Phillips, C. (2015). The role of structural prediction in rapid syntactic analysis. Brain and language, 140, 9-21.

Lee, M. W. (2004). Another look at the role of empty categories in sentence processing

(and grammar). Journal of Psycholinguistic Research, 33, 51-73.

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

Lowder, M. W., & Gordon, P. C. (2014). Effects of animacy and noun-phrase relatedness on the processing of complex sentences. Memory & cognition, 42, 794-805.

Lowder, M. W., & Gordon, P. C. (2015). Natural forces as agents: Reconceptualizing the animate–inanimate distinction. Cognition, 136, 85-90.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996).

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. Journal of memory and language, 47(1), 50-68.

McElree, B., & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(2), 432.

Miyamoto, E. T. (2002). Case markers as clause boundary inducers in Japanese. Journal of psycholinguistic research, 31, 307-347.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. Frontiers in psychology, 6, 384.

Packard, J. L., Ye, Z., & Zhou, X. (2010). Filler-gap processing in Mandarin relative clauses: Evidence from event-related potentials. In Processing and producing head-final structures (pp. 219-240). Dordrecht: Springer Netherlands.

Parker, D. (2017). Processing multiple gap dependencies: Forewarned is forearmed. Journal of Memory and Language, 97, 175-186.

Phillips, C., & Lau, E. (2004). Foundational issues. Journal of Linguistics, 40(3), 571-591.

Pickering, M. J., & Traxler, M. J. (2003). Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. Language and Cognitive Processes, 18(4), 469-503. Staub, 2010;

Rayner, K., & Pollatsek, A. (1989). The psychology of reading.

Saito, M., Lin, T. H. J., & Murasugi, K. (2008). N'-ellipsis and the structure of noun phrases in Chinese and Japanese. Journal of East Asian Linguistics, 17(3), 247-271.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. The Oxford handbook of psycholinguistics, 327, 342.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. Language and cognitive processes, 1(3), 227-245.

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

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. Journal of Memory and Language, 35(3), 454-475.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: evidence from eye movements. Journal of Memory and Language, 47, 69-90.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005) Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and language*, 53(2), 204-224.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PloS one*, 8(10, e77006.

Wanner, H. E., & Maratsos, M. P. (1978). An ATN approach to comprehension. In Linguistic theory and psychological reality. Mit Press.

Wu, F., Kaiser, E., & Andersen, E. (2009). The effect of classifiers in predicting Chinese relative clauses. In the Proceedings of the Western Conference on Linguistics (WECOL). Davis: University of California. Proceedings online: http://wecol. ucdavis. edu.

Wu, F. (2011). Frequency issues of classifier configurations for processing Mandarin object-extracted relative clauses: A corpus study.

Wu, F., Luo, Y., & Zhou, X. (2014). Building Chinese relative clause structures with lexical and syntactic cues: Evidence from visual world eye-tracking and reading times. Language, Cognition and Neuroscience, 29(10), 1205-1226.

Wu, F., Kaiser, E., & Vasishth, S. (2018). Effects of early cues on the processing of Chinese relative clauses: Evidence for experience-based theories. Cognitive science,

42, 1101-1133.

Yoshida, M. (2006). Constraints and mechanisms in long-distance dependency formation (Doctoral dissertation).

Zhang, N. N. (2013). Classifier structures in mandarin Chinese (Vol. 263). Walter de Gruyter.

Zhang, Y., Zhang, J., & 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.

Zhou, X., Jiang, X., Ye, Z., Zhang, Y., Lou, K., & Zhan, W. (2010). Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study. Neuropsychologia, 48(6), 1551-1562.

Zhu, X. (2019). Word predictability and word class modulate the predictive N1 in children and adults. Developmental Cognitive Neuroscience, 37, 100614.

Article 3

Negative islands do not block filled-gap effects

Zirui Huang and E. Matthew Husband

University of Oxford

Mailing address:
E. Matthew Husband
St. Hugh's College
St. Margaret's Rd.
Oxford OX2 6LE
United Kingdom

Email:
zirui.huang@ling-phil.ox.ac.uk (Huang),
matthew.husband@ling-phil.ox.ac.uk (Husband)

Running head: NEGATIVE ISLANDS & FILLED-GAPS

Word Count: 9520

Abstract[1]

Research has found that the parser respects strong island constraints when actively positing gaps for displaced phrases, reflecting the parser's rapid use of syntactic constraints to avoid positing tempting but illicit dependencies in real-time. Long-distance dependencies, however, are also governed by semantic/pragmatic constraints. Negative islands, a type of weak island, selectively constrain certain wh-dependencies that violate Dayal's (1996) maximal informativity presupposition on questions, i.e., that the answer set contains a true answer entailing all the other true ones (Fox & Hackl, 2007; Abrusán, 2011). Whether the parser can use a presuppositional constraint like negative islands in real-time to avoid positing these types of illicit dependencies is unclear. An offline acceptability task, a self-paced reading, and an eye-tracking study were conducted to examine whether weak negative islands are as effective as strong wh-islands at blocking illicit gaps. Our results suggest that although comprehenders are aware of negative island constraints offline, the parser is unable to use them to block real-time active dependency formation. This asymmetry suggests that the effects of weak (semantic/pragmatic) islands take time to emerge, unlike strong (syntactic) islands which are more immediate.

Key words: negative islands; semantic processing; eye-tracking; self-paced reading.

[1] Abstracts of this study has been submitted for academic conferences including 2nd Annual Conference on Experiments in Language Meanings; 29th Architectures and Mechanisms for Language Processing; 10th biennial meeting of Experimental Pragmatics.

Introduction

Whether the mechanisms of language processing respect grammatical constraints has been a long-standing issue in developing a comprehensive theory of linguistic performance. Offline acceptability judgments have been the main empirical measure for linguists in the development of theories of linguistic competence, while psycholinguists have focused on online (real-time) effects that track the dynamic processes of language processing. The gap between evidence from online and offline observations raises questions concerning the grammar-parser relation: to what extent does our grammatical knowledge guide real-time sentence processing mechanisms?

A classic view of the grammar-parser distinction found in Bever (1970) states that grammatical rules function as a backup system applied after heuristic parsing has provided an initial analysis. This dual-analyzer account is encapsulated in the slogan “we understand everything twice” (Townsend & Bever 2001), which claims that sentences are first processed by a “quick and dirty” parser, and then grammatical rules are applied to the interpretation of sentences when needed. The heuristic nature of the parser has been demonstrated in many studies in terms of issues concerning ambiguity resolution, where the parser seems to not entertain certain possible analyses (e.g., Grant et al., 2020; Swets et al., 2008; Traxler et al., 2009), and grammatical illusions like agreement attractions, illusory NPI licensing, comparative illusions, etc., where the parser appears to adopt an ungrammatical analysis (e.g., Bock and Miller 1991; Clifton et al. 1999; Pearlmutter et al. 1999). Sentence processing theories and models, such as “good-

enough" theories (Ferreira, Bailey & Ferraro 2002; Ferreira & Patson 2007; Karimi & Ferreira 2016) have shown that the initial analysis conducted by the heuristic parser disregards grammatical rules when processing such complex or ambiguous constructions and thus only incomplete understandings of grammatical structures are achieved. Dual-analyzer accounts are also empirically supported by the misalignments between results of offline judgments and online data found in cases involving "grammatical illusions", where comprehenders demonstrate sensitivity to ungrammatical sentences in offline data while failing to detect such grammatical errors in online sentence comprehension, such as garden path phenomenon and revision failures (e.g., Bever, 1970; Sturt, 2007), center embeddings and processing overload (e.g., Frazier 1985; Gimenes et al. 2009; Trotzke et al, 2014), and agreement illusions (e.g., Bock and Miller 1991; Clifton et al. 1999; Pearlmutter et al. 1999).

On the other hand, a tighter conception of the grammar-parser relationship has also been widely considered (Phillips, 1996; 2012; Lewis and Phillips, 2015). Under the view of a single-analyzer, the notion of "parser" and "grammar" describe the same language system from different perspectives: grammar represents an idealized model of real-time language comprehension with unlimited cognitive resources (Phillips, 1996). Apparent differences between the grammar and parser arise from constraints related to limits on attentional resources and a noisy memory architecture. There is substantial evidence indicating that real-time processing can be grammatically precise during sentence processing, for example, the parser's sensitivity to island constraints (e.g., Stowe 1986;

Phillips 2006; Traxler and Pickering 1996; Wagers et al. 2009; Yoshida et al. 2004; Omaki and Schulz 2011); successful avoidance of ungrammatical backward anaphora (Cowart and Cairns 1987; Kazanina et al. 2007; Aoshima et al. 2009) and parser's ability to ignore illicit antecedents for the interpretations of reflexive pronouns (e.g., Nicol and Swinney 1989; Dillon et al. 2013; Sturt 2003; Badecker and Straub 2002; Clackson et al. 2011; Xiang et al. 2009).

The current study aims to address the issue of grammar-parser distinction by examining the real-time status of negative island constraints on language processing. Island constraints have been found to grammatically constrain unbounded linguistic dependencies (Ross, 1967). The effects of island constraints have been successfully tracked by both offline and online measures (e.g., Stowe, 1986; Frazier, 1987; see detailed description in the later section), indicating a unified relation between real-time parser and grammar knowledge. For example, the filled-gap effect (Stowe, 1986) has been used as a typical observation in real-time measures for dependency formation and the effect of island constraints. It demonstrates that the parser actively posits gaps for wh-phrases in grammatical positions and filled-gap effects emerged when gaps are grammatically licensed but not when they are grammatically inaccessible, e.g., inside a Complex NP Island. It reflects the parser's rapid use of syntactic grammatical constraints to avoid positing illicit dependencies in real-time. However, the vast majority of studies focusing on a parser's grammatical sensitivity tested strong island constraints. Negative islands, unlike strong syntactic islands, are a type of weak island that arises from

semantic/pragmatic considerations. It has been proposed that wh-dependencies in degree questions that move across negation often violate Dayal's (1996) maximal informativity presupposition on questions, i.e., that the answer set contains a true answer entailing all the other true ones (Fox & Hackl, 2006; Abrusán, 2011). It is illustrated by the acceptability differences between (1a) and (1b).

(1) a. How tall is John?

b. * How tall isn't John?

Negative degree questions such as (1b) are judged to be unacceptable because the presupposition of a maximally informative answer requires a minimal height interval that does not contain John's actual height. However, such an interval does not exist because the true answer set contains two mutually exclusive subsets that do not entail one another, i.e., all intervals below John's height, $(0, \text{height}_{\text{John}})$, and all intervals above John's height, $(\text{height}_{\text{John}}, \infty)$. Compared to syntactic islands, the calculation required to implement this negative island constraint may be more complex than those required to recognize strong syntactic islands. It may take the parser more resources and time to use such a violation to block dependency formations. Thus, negative islands provide a testing ground to examine whether this presuppositional grammatical constraint can also be utilized by a parser in real-time sentence processing. Evidence about the processing of negative islands could potentially help us to gain a better understanding of grammar-

parser relations by establishing whether the parser considered such constraints on grammatical well-formedness. To our best knowledge, the real-time status of negative island processing is currently unknown. It is important to understand how a negative island is recognized in real-time sentence comprehension and how different it is from other syntactic island processing. Furthermore, it might shed light on whether the parser can demonstrate greater capacity to handle more complex and subtle grammar constraints. In this study, we examined whether negative islands are as effective as syntactic islands at blocking illicit gaps in real time.

In the following part of the introduction, we first overview prior works on the real-time processing of island constraints in general, and then we specifically introduce studies that have been done about negative island constraints. In the next section, we present our experiment design to investigate the real-time processing status of negative island and show our analysis and results. Then discussion and conclusion are given in the final section.

Real-time processing of island constraints

Island constraints are one of the most well-studied grammatical constraints in psycholinguistics. They place limits on the grammaticality of unbounded dependency formation. For example, the formation of a dependency relation caused by WH-movement is unbound in (2a), however the completion of such a dependency is restricted when the canonical gap for the wh-word is located in a relative clause as in (8b). When

the proposed gap is located inside an island domain, a dependency relation that goes across the island boundary is prohibited.

(2) a. Who$_i$ did you hope that the candidate said that he admired __$_i$?

b. *Who$_i$ did the candidate read a book that praised __$_i$?

Many studies have proposed that such islands have a syntactic nature. Ross (1967) proposed several independent grammatical constraints on dependencies, separately mapping to different types of islands: e.g. Complex-Noun Phrase Constraints, Coordinate Structure Constraints, Left Branch Condition, and Sentential Subject Constraints. Chomsky (1973) merged these constraints and proposed the Principle of Subjacency, which prohibits dependency relations over two or more bounding nodes. Later Huang (1982) proposed the Condition on Extraction Domains, which states that dependency relations cannot be established from non-complements, i.e., subjects and adjuncts. Huang's proposal was incorporated into the later version of the barriers approach in the principles-and-parameters framework (Chomsky, 1986).

This grammatical knowledge is also reflected in the real-time processing of filler-gap dependencies. Psycholinguists have established an active gap filling mechanism for the parser when processing filler-gap dependencies (Fodor, 1978; Crain and Fodor, 1985; Stowe, 1986; Frazier, 1987), which means that instead of waiting for unambiguous

bottom-up identification of a gap that completes a dependency, the parser actively posits a gap site at the early as possible, potentially to ease the working memory burden of carrying an unintegrated filler. This proposal is supported by the observation of filled-gap effects (Stowe, 1986). Such an effect shows up when the parser predicts a gap site in advance but then finds the postulated grammatical gap is taken up by a linguistic element, requiring it to reanalyze the dependency. Longer reading time at the filled element is considered to reflect processing penalties for incorrect expectations. Stowe (1986) found inflated reading times at “us” in (3a), where the filler “who” triggers an active expectation of a gap at the grammatical position taken up by “us”, relative to “us” in (3b), where no dependency relation is expected to be completed.

(3) a. My brother wanted to know who Ruth will bring us home to at Christmas.

b. My brother wanted to know if Ruth will bring us home to mom at Christmas.

Relevant to the relationship between grammar and parser is the observation that island constraints modulate the filled-gap effect during sentence processing. No filled-gap effect is observed inside of an island domain, suggesting that the parser is aware of the fact that extraction out of an island is not grammatically permitted. For example, in Stowe’s study, no filled-gap effect was found at the noun phrase “Greg’s older brother” in (4a) as the NP “Greg's”, as part of a complex subject NP, compared with (4b) where no dependency is expected at all. This finding suggests that although the parser actively predicts gaps due to extraction, it is also sensitive to the grammatical information that no

such gap can occur inside this island domain. Therefore, the presence and absence of filled-gap effect paradigms have been commonly used as a real-time processing indicator in studies involving long-distance dependencies and island constraints (e.g., Frazier and Clifton, 1989; Pickering, Barton, and Shillcock 1994; Fujita and Cunnings, 2020).

(4) a. The teacher asked what the team laughed about Greg's older brother fumbling.

b. The teacher asked what the silly story about Greg's older brother was supposed to mean.

Similar results also come from other studies making use of mismatch plausibility effects. For example, an eye-tracking while reading study by Traxler and Pickering (1996), manipulated the plausibility between a displaced filler such as "the book" or "the city" and a verb complement, as in "they talked about the book/city that the author wrote __…", and found longer reading time at the verb when the filler was a semantically implausible object for the verb (city-wrote) compared with when the filler was a semantically plausible object (book-wrote), which is consistent with the active gap filling strategy. More crucially, no plausibility mismatch effect was found on the verb "wrote" was embedded in a wh-island, as in "the book/city that the author who wrote __ …" Here, the potential gap is not grammatical licensed for the filler-gap dependency due to wh-island constraints. The fact that the filler-gap dependency formation is sensitive to island constraints suggests a rapid use of grammatical cues and knowledge in filler-gap dependency formation.

These island effects as mentioned above provide syntactic environments where extraction is categorically forbidden, and they are referred to as strong islands. Violations of strong island constraints result in clear ungrammaticality. In contrast, weak islands are environments that restrict extraction but are not absolute constraints that block all extractions as strong islands do. Weak islands tend to impose selective constraints, depending on specific properties of the extracted element, normally with additional semantic requirements, such as why, how, or degree/quantity expressions. The reason for the selectivity is that weak island constraints often arise from contextual or semantic/pragmatic considerations that involve the validation with real-world knowledge beyond literal inputs. For example, certain environments, particularly those that involve presuppositional or factive verbs (e.g., regret, realize), create weak island effects. For example, (5a) is generally judged as unacceptable because the extraction is not specific or definite. In the case of a more specific extraction like (5b), it becomes more acceptable. The extraction of “how” is more problematic compared to the extraction of “what”, indicating that presuppositional islands are sensitive to the specificity or definiteness of the extracted element. Negative islands are another type of weak island which arises when negation in the sentence creates difficulties for certain types of extraction, especially for quantificational elements or adjuncts. For example, (5c) compares to (5d), the extraction of degree questions (e.g., *how many people*) is blocked in the presence of negation, while the extraction of individual arguments (e.g., *who*) is allowed. This indicates that a weak island created by negation affects only specific types of

dependencies, particularly those involving manners and degrees.

(5) a. * How do you regret that you didn't buy __?

b. What do you regret that you didn't buy __?

c. * How many people didn't John invite __?"

d. Who didn't John invite __?"

The fact that weak islands, no matter which type, only provide selective constraint, may suggest that mechanisms for the weak island constraint are different from strong islands. In fact, it has been a great challenge for any account that proposed that weak island constraint, such as a factive island in (5a), or negative island in (5b), is caused by some syntactic operation violation as proposed for strong islands (e.g., Relativized Minimality by Rizzi, 1990). Moreover, despite various semantic explanations proposed for weak islands, there is currently no coherent account that can comprehensively explain all types of weak island phenomena. Szabolcsi and Zwarts (1993) developed a unified, algebraic-semantic-based approach to explain the weak island intervention effects. Their account hinges on the idea that certain operators, such as negation or quantifiers, alter the semantic environment that disrupts the establishment of a dependency relation to form across them. They argue for the critical role that semantic factors can play in shaping whether or not long-distance extraction is viable. However, this account does not touch upon all types of weak islands like factive islands. Special

attention for weak islands has been on the negative islands caused by degree or manners questions. For example, Fox and Hackl (2007) proposed a semantic account for negative degree islands constraint that when degree operators are involved, the wh-dependencies require a semantically coherent mapping from the filler to the gap. When negation comes into play, it creates a configuration in which the degree operator cannot be interpreted properly, yielding a negative island effect. However, it remains unclear whether these accounts can be extended to all weak island constraints. Rullmann (1995) explicitly expressed skepticism that a unified account for different types of weak islands is possible.

Consequently, weak islands exhibit more subtle effects in real-time processing compared with strong island constraints introduced earlier. In Kluender and Kutas (1993), they show unbounded dependencies elicit distinct ERP signatures (particularly a P600) when disrupted. Their results also revealed that these weak island violations trigger measurable neural responses indicative of syntactic difficulty. Although weak island effects were somewhat less pronounced than those observed for strong island violations (e.g., relative clauses, complex NPs), the ERP profile still showed that the parser was sensitive to the presence of the embedded wh-phrase. The amplitude and scalp distribution of the P600 suggested that, even in “weaker” contexts, the parser recognized these as illicit extraction sites. A more recent study by Villata et al. (2020), employing acceptability judgment tasks and maze-based self-paced reading, provides evidence that weak islands may influence active dependency formation in ways comparable to

strong islands. Their investigation examined a range of different island types including both weak islands (e.g., whether-islands) and strong islands (e.g., adjunct islands and complex NP islands). Furthermore, eye-tracking experiments conducted by Cokal and Sturt (2022) show that one type of weak island – the *whether* island exhibits a similar blocking effect for the filled-gap effect typically observed in strong islands, compared with non-island condition, suggesting that the parser has real-time sensitivity for both strong and weak islands. However, given the diversified nature of different types of weak islands, results for one type of weak island do not speak for the weak island as one unified category. Different types of weak islands might be originated from different levels of grammatical considerations. Whether-islands are still considered to be syntactic in nature with clear syntactic configurational boundaries, while on the other hand, negative island, another type of weak islands, arises from semantic/pragmatic considerations. The negative island caused by degree questions or manner questions is generally considered to be a pragmatic island constraint caused by presupposition violation. Thus It is important to understand how a negative island is recognized in real-time sentence comprehension and how different it is from other syntactic island constraints. By comparing the real-time status of the negative island with a strong WH-island, we aim to investigate how syntactic cues and presuppositional cues differ in blocking the filler-gap dependency formation, rather than simply contrasting strong islands and weak islands are two distinctive categories.

Negative island constraints

The term negative island (Ross, 1984) has been used to refer to the unacceptability of constructions when attempting to form a wh-dependency across negation. In contrast with strong islands, where all types of extraction are not allowed, negative islands are a type of weak island that only selectively rejects extraction in certain constructions. As shown in (6a-d), extracting a which-question over negation (6b) is generally judged acceptable, but extracting a degree question over negation (6d) is generally unacceptable.

(6) a. Which man did John invite to the party?

b. Which man didn't John invite to the party?

c. How many children did the dog scare?

d. * How many children didn't the dog scare?

However, this negative island violation can be obviated by introducing context, adding extra modal verbs or attitude verbs. For example, if comprehenders are introduced to a situation that there are 12 children relevant in the context, then (6d) would become more acceptable. Also as noted by Fox and Hackl (2007), negative island violation can be significantly obviated by introducing an existential modal within the scope of the negation. See (7a-b) as an example:

(7) a. * How many children don't you have?

b. How many children are you not allowed to have?

Various accounts have been provided for negative islands, invoking different levels of representations. A syntactic account has been proposed to explain negative islands by introducing negation as a potential antecedent governor that governs wh-movement (Rizzi, 1990). Rizzi's syntactic account divides the wh-movement types into referential and non-referential expressions. A wh-expression assigned with a referential theta role can be extracted while a wh-expression without a referential theta role, e.g., degree, manner and measure expressions, cannot undergo extraction. Thus, negation as a potential antecedent governor rejects extraction of a non-referential wh-phrase but permits referential wh-extraction. However, such a purely syntactic account cannot explain the obviation effects observed in (6) and (7), that a syntactic violation can be ameliorated just by manipulating the context or extra-structural elements. An alternative to the syntactic approach adopts a semantic/pragmatic analysis (Kluender and Gieselman 2013; Rizzi 2003, 2004). Szabolcsi and Zwarts (1993) suggest that the semantic properties of wh-expressions determine whether they can be legitimately extracted. Which-phrases, for example, demonstrate the characteristic of being an unordered element in a set of discrete individuals, while other wh-phrases, like manner and degree expressions, denote a domain that is not individuated. They propose that negation permits extractions to happen with discrete individuals but rejects extractions

with unindividuated phrases. Apart from focusing on the properties of wh-expressions, a semantic/pragmatic approach to negative island constraints pays attention to the pragmatic principles of questions. Questions are subject to a presupposition of maximal informativity (Dayal, 1996), which states that all questions require one unique answer that is maximally informative, i.e., the answer set to a question must contain one true answer that entails all the other true answers (Fox & Hackl, 2006; Abrusán, 2011). For example, (8a) meets this presupposition in the way that there is a definite number of John's height, thus a most informative answer can be provided for this question. Degree questions like (8d) violate this presupposition. Because of the density of height, the true answer sets to this question are open intervals, resulting in infinite true answers (say John is 178cm high, then 179cm, 180cm, 185cm, 200cm… are all true answers) but not one maximal true answer that entails all the other true answers. Comprehenders with world knowledge understand that it is impossible to find the maximal informative answer to this question, thus it is judged to be unacceptable to form such a wh-extraction.

(8) a. How tall is John?

b. * How tall isn't John?

The presuppositional constraint accounts can also explain why modal verbs are able to rescue a negative degree question, which cannot be explained based on the accounts focusing on the properties of wh-expressions. Model verbs alleviate the presupposition

requiring a uniquely and maximally informative answer by shifting the interpretation of the degree question to a possibility space, allowing for a broader range of acceptable answers. as shown in the example (9a) and (9b).

(9) a. * How slow didn't he drive?

b. * How slow shouldn't he be driving?

So far, the exploration of the negative islands has mostly been on providing accounts for its nature in the theoretical linguistics field with evidence from off-line grammatical judgment. However, work on the real-time status of negative island constraint is still waiting to be filled. The comparison between a presuppositional island and a syntactic island in terms of real-time processing sensitivity has, to my best knowledge, not been investigated. Thus it still remains unclear whether a negative island constraint can be rapidly utilized by a real-time parser in sentence processing, or it is temporally ignored by the parser and is revised in the later stage. By addressing this issue, we can gain a better understanding of how the parser recognizes the properties that define an island domain, and also the observation can further contribute to the discussion of the relationship between static grammar knowledge and dynamic mental parsing mechanism.

This study

The purpose of this study is to detect the time course of negative island constraint in real-time sentence processing. One offline acceptability judgment experiment and two real-time experiments, including one self-paced reading study and one eye-tracking while reading study, were conducted to see if the reading pattern of negative islands is more in line with strong wh-islands or shows more similarity with null island sentences. Given the fact that our mental parser is highly incremental, we are able to observe some temporary reading difficulties/facilitations which indicate the parser's analysis preferences. As reviewed in the earlier section, filled-gap effects have a long history of being used to study real-time long-distance dependency formation. In this design, we also target at the classic filled-gap effect that is typically observed in the comparison between strong wh-island and no-island sentences, and add sentences (temporarily) containing negative islands. See (10a-c) for example, for an incremental parser, a filled-gap effect is expected to show up at "famous" in (10a) compared with (10b), because there is no overt configurational mark indicating the occurrence of an embedded island domain and the active gap filling strategy would prompt the parser to posit a gap at the earliest possible position to host the wh-phrase "how tall". Then a filled-gap effect would show up as the postulated grammatical position is taken up by "famous". While (10b) provides an overt configurational boundary ("who was") for a strong wh-island, due to the parser's sensitivity to strong island constraint, active gap filling is blocked, and no such filled-gap effect will be observed at "famous" in (10b). Then we focus on whether a

degree wh-filler in combination with a negation can be recognized as a cue to block the active gap filling in (10c). If there is a filled-gap effect at “famous” in (10c), it might suggest that seeing “how tall” and then a negation does not ring a bell for the parser to consider that it violates a presupposition of question formation, i.e., the parser cannot make use of negative island constraint in real-time. On the other hand, if there is no filled-gap effect in (10c), it might suggest that the parser is aware that an extraction over negation is not semantically plausible in this context, i.e., negative island might be as effective as a strong wh-island.

(10)

a. How tall did Mary think the girl hoped to be **famous**…… (No-Island)

b. How tall did Mary think the girl who was hoped to be famous…… (Wh-Island)

c. How tall did Mary think the girl hoped not to be famous…… (Neg-Island)

Experiment 1: Offline acceptability judgment

We first conducted an offline acceptability judgment study, to examine the general acceptability of different sentence structures and to what extent different types of island constraints affect grammaticality judgment.

Participants.

51 participants were recruited through an online participant recruiting platform Prolific. All of them were claimed to be English native speakers. They were paid £2.75 for their participation.

Material

We manipulated POLARITY (Positive, Negative) and STRUCTURE (No, Reduced relative clause, Unreduced relative clause), creating six conditions for experimental items. An example set of experiment items is presented in (11a-f). We used adjectival degree questions, e.g., “how tall”, to make sure that the presuppositions for our experimental items are not contextually dependent as noted for example (6). Experimental items were counterbalanced and evenly distributed into 6 lists using a Latin square design, ensuring that participants rated through each condition but never saw related lexicalizations within and across conditions. Besides experimental items, we also adopted acceptability judgment materials from Sprouse (2011) which provide sentences ranging from extremely ungrammatical to perfectly grammatical. These items were used as the “anchoring” items to balance the grammaticality judgment range of this study. In total 72 items (24 experimental items, and other anchoring items and filler items) in each list were pseudo-randomized to make sure that items in the same condition never appeared consecutively. Participants were randomly assigned a list and the total time to finish rating a list is around 5 minutes.

(11) Example item in Experiment 1 acceptability judgments

a. How tall did Mary think the girl hoped to be? (NoRC, Positive)

b. How tall did Mary think the girl hoped not to be? (NoRC, Negative)

c. How tall did Mary think the girl hoped to be to be famous by her parents was? (RRC, Positive)

d. How tall did Mary think the girl hoped not to be to be famous by her parents was? (RRC, Negative)

e. How tall did Mary think the girl who was hoped to be to be famous by her parents was? (URC, Positive)

f. How tall did Mary think the girl who was hoped not to be to be famous by her parents was? (URC, Negative)

Procedure

This study is presented as an online survey, so participants used a web browser through in their own place and pace to complete the task. They were first given instructions stating that they should indicate their acceptability of the sentence they read by giving a number from 1 to 7, which 1 indicated that they considered the sentence as very unacceptable and 7 indicated that the sentence was perfectly acceptable.

Results

We ran linear mixed effects models with items and participants introduced as random factors, and polarity and sentence type were fixed factors. Then we obtained pairwise comparison results by using the Emmeans package in R.

Overall there is a main effect of polarity (Est. = 0.34, t= 8.56). As for the effect of structure, main effects were found on contrasts of no island structure v.s. Relative clause

island structure (Est. = 0.74, t=13.37), and of reduced relative clause structure v.s. unreduced relative clause structure (Est. = 0.32, t= 6.66). Besides, an interaction between polarity and sentence structure is also significant when contrasting no island structure with relative clause island structures (Est. = 0.38, t=6.87) but no interaction is found in the contrast of reduced relative clause structure v.s. unreduced relative clause structure (Est. = 0.02, t= 0.37). Table 1a and 1b present the model summary and further pairwise comparisons result of the polarity effect within sentence structures, and Figure 1 shows the averaged ratings across all participants.

	Est.	***t***	***p***
Polarity	0.34	8.56	<0.001
No RC vs RCs	0.74	13.37	<0.001
RRC vs URC	0.32	6.66	<0.001
Polarity:No RC vs RCs	0.38	6.87	<0.001
Polarity:RRC vs URC	0.02	0.37	0.71

Table 1: Model Summary for Experiment 1

Positive – Negative	***Est.***	***t***	***p***
No RC	1.48	10.56	<.001
Reduced RC	0.27	1.94	.053
Unreduced RC	0.38	2.46	.014

Table 2: Effect of polarity within sentence structures for Experiment 1

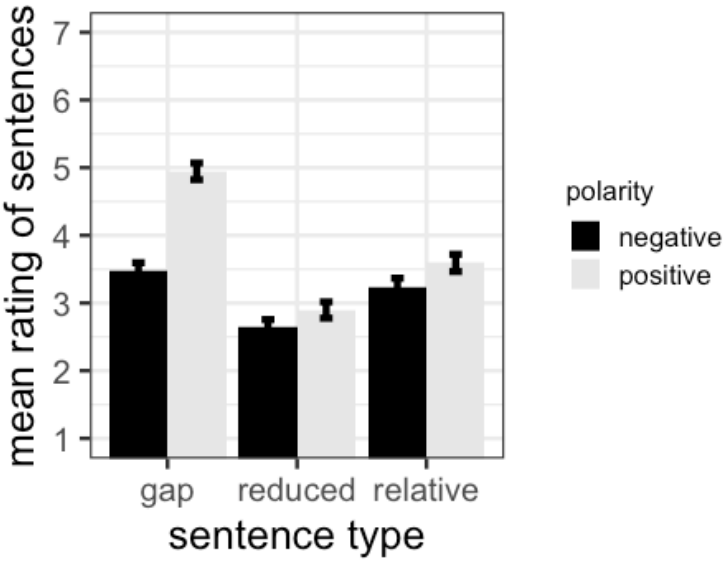


Figure 1. Acceptability judgment ratings

Discussion

To start with, the main effect of polarity demonstrates that participants in general rated degree questions with negation lower than degree questions without negation, suggesting that negation is in general dispreferred to co-occur with degree questions. Then the effect of sentence structure shows that participants dispreferred complex and/or temporarily ambiguous sentences. Sentences with embedded relative clauses were rated lower than simple main clauses, and reduced relative clauses were even lower than unreduced relative clauses. Reduced relative clauses might cause temporal ambiguity and lead participants to a garden path of simple main clause structure. Sentences were disambiguated later at the prepositional phrase region (“by her parents” as in the example stimuli) or even at the main verb (“was”). The complexity of relative clause structure and garden path phenomenon might affect the participants to give a lower acceptability score. More relevant to this study, we found an interaction between

polarity and sentence structure that participants gave much lower scores for negations in the main clause than negations embedded in the relative clause. The much lower ratings in negative main clauses compared with positive main clauses suggest that participants realized the violation of maximal informativity and considered it unacceptable. The narrowed difference in ratings in relative clauses suggests that relative clause island effectively blocks the scope of negation and participants understood that the negation did not affect the degree question in the main clause, rescuing the acceptability of negation with relative clauses. More importantly, within the relative clause structures, there is no interaction of polarity effect and reduced/unreduced relative clauses, showing the effect of negation (reduced relative clause) is more in line with the effect of wh-filler (unreduced relative clause). It might suggest that participants were able to use negative island constraint as well as strong island constraints offline.

Experiment 2: Self-paced reading

The acceptability judgment results from Experiment 1 set up this self-paced reading experiment to use online filled-gap effects to investigate whether the parser posits illicit gaps inside negative islands compared to wh-islands.

Participants

66 self-reported native English speakers were recruited via an online platform Prolific. Three participants were excluded from further data analysis due to poor performance with an accuracy rate lower than 60%. All participants were paid £6.75 as compensation for their time participating in this experiment.

Material

We created 24 items manipulating ISLAND type[2]: no island, negative island and wh-island (strong island) to examine whether the parser actively posits a (temporary) gap inside islands. An example set of experiment items is presented in (12a-c). In addition, 48 filler items were used and combined with the experimental items. None of the filler items contain filler-gap dependencies or negation but they were of comparable structural complexity with anaphoric pronoun resolution problems and various syntactic ambiguities.

(12) Example items in Experiment 2: self-paced reading

a. How tall did Mary think the girl hoped to be famous by her parents was before she went to college? (No-Island)

b. How tall did Mary think the girl hoped not to be famous by her parents was before she went to college? (Neg-Island)

c. How tall did Mary think the girl who was hoped to be famous by her parents was before she went to college? (Wh-Island)

[2] Our initial design was a 2×2 factorial design manipulating Polarity (Negative vs. Positive) and Island type (Present vs. Reduced). However, during piloting and initial data analysis, we found that the combination of a negation and relative clause (e.g., How tall did Mary think the girl who was hoped not to be famous by her parents was before she went to college?) resulted in sentences that were syntactically and semantically more complex than the other conditions. Comprehenders in the pilot study appeared to triage these sentences, treating them as anomalous and in practice dropping out of normal, incremental language comprehension. This asymmetry made it difficult to construct natural and comparable items across all cells of the 2×2 design.

To ensure experimental balance and interpretability, we therefore opted for a three-way factorial design (No-island vs. WH-island vs. Neg-island). This approach preserved the central contrast, which is to test whether sentential negation behaves as an island boundary, while avoiding potential confounds introduced by unnatural or structurally unbalanced materials. We believe this adjustment yields clearer results regarding the effect of negation on filler–gap dependency formation.

Procedure

The experiment was built upon a web-based platform IbexFarm (Drummond, 2013). Participants got access to this experiment via a link sent out by the recruitment platform Prolific. Sentences in this experiment were presented in a word-by-word fashion. Only one word was displayed on the screen at a time and participants pressed the spacebar to proceed to the next word. Reaction times for pressing the spacebar to display the next word were recorded to measure reading times for each word. Half of all sentences were followed by a yes/no comprehensive question as attention checks to make sure participants concentrate on this online task.

Upon entering the experiment surface, participants were presented with informed consent and instructions about the experiment. They are told to read the words at a natural pace and answer the comprehension questions as accurately as possible. After participants agreed to proceed and read through the instructions, five test trials were presented to make sure they understood the procedure. Then another action of pressing the spacebar needs to be taken to start the experiment. The total time for completing a study was roughly 10-15 minutes.

Results

Data were analyzed using linear mixed effect models in the lme4 packages in R. As mentioned earlier in the participant section, Data from three participants were excluded

due to poor accuracy rate to comprehension questions (lower than 60%). Reading times that fell outside 2 standard deviations for the overall mean reading time for each word were also excluded from further analysis. The critical word where the filled-gap effect should emerge is the adjective embedded in the relative clause (“famous” as in the example stimuli). The prepositional phrase following the adjective (“by her parents”) is also controlled as spillover regions. “By” is the spillover region 1 and “her” is the spillover region 2. We made comparisons between the no-island condition and wh-island condition, to observe the classic filled-gap effect with strong island constraint, and also between no-island and negative island condition, to see if there is any filled-gap effect for negative island constraint.

The results are plotted in Figure 2 and the model summaries are presented in Table 3. At the critical adjective region, we observed an expected filled-gap effect in the comparison between no-island and wh-island (Est. = 45.72, t = 2.05, p = 0.04), showing that wh-island can effectively block a filler-gap dependency formation. The filled-gap effect continued in the first spillover region (Est. = 41.37, t = 2.632, p = 0.008) and the second spillover region (Est. = 22.21, t = 2.55, p = 0.01). However, no effect was found between no-islands and negative islands at the critical adjective region (p = 0.35) or any spillover region (first spillover region: p = 0.61; second spillover region: p = 0.95). Numerically, there seems to be a trend of faster recovery from the filled-gap effect at the spillover region in negative island condition, but there is no statistical significance.

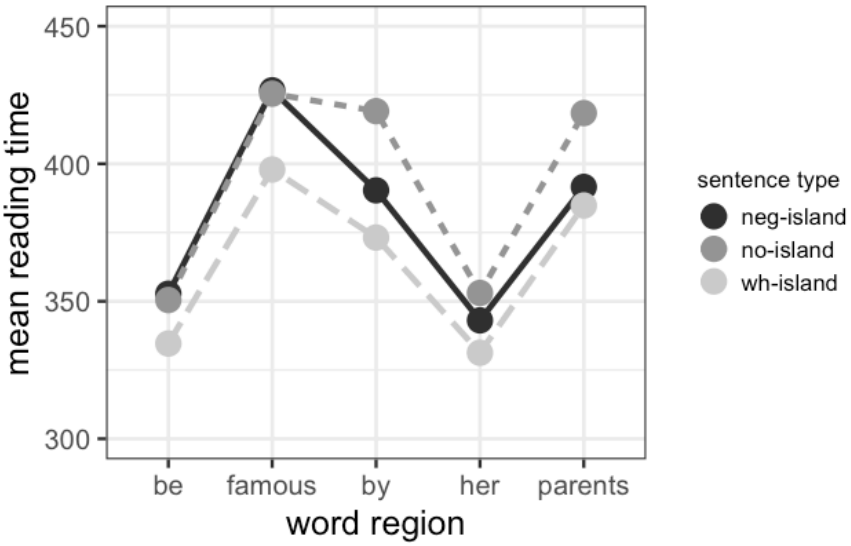


Figure 2. Mean self-paced reading time by word regions

Discussion

The results suggest that such a semantic island constraint might take time to emerge, unlike a syntactic island, which is more immediate. To probe further into the negative island constraint in the processing of dependency completion, we consider eye-tracking while reading as a suitable method to demonstrate the time course of negative island effects.

Experiment 3: Eye-tracking while reading

We used the same experimental items in Experiment 2 to conduct an eye-tracking study. Eye-tracking method allows a more naturalistic paradigm for reading comprehension tasks compared with self-paced reading in which participants are forced to read forward region-by-region, and also provides more reading time measurements to

investigate the time-course of the negative island effect that might be collapsed in self-paced reading paradigm.

Participants

30 self-reported native English speakers were recruited in Oxford. All of them reported normal or corrected-to-normal vision. Four participants were excluded from further data analysis due to poor performance with an accuracy rate lower than 75%. All participants were paid £5 as compensation for their time participating in this experiment.

Material

The material used for the Experiment 3 eye-tracking study was identical to the items used in the Experiment 2 self-paced reading study. In total 72 items (24 experimental items and 48 filler items) were evenly divided into 2 experimental blocks. There was also an additional practice block before the experimental blocks for participants to get used to the procedure.

Procedure

Eye movements were recorded with an EyeLink 1000 plus system (SR Research, Toronto, Ontario, Canada) at a sampling rate of 1,000 Hz. Calibration and validation were performed at the beginning of each block and throughout the experiment as needed. A drift correction was also performed at the beginning of each trial. The participants were given instructions prior to the start of the experiment. They were asked to read the sentence on the screen and to fixate on a black dot at the top-right corner area of the screen when they finished reading, then a true/false question related to the previous

sentence was displayed and the participants should answer the question according to their understanding of the previous sentence by pressing a button. Half of the trials were followed by comprehensive questions and the participants were told to read the sentences carefully and make sure to answer the true/false questions correctly. They then first enter a practice block to get familiar with the procedure. After this practice session, they would go through two experimental blocks. Participants were given time to take a break between the two blocks. They were also told that they could rest their eyes whenever they felt like during the experiment, and then calibration and validation would be performed again when they were ready.

Data analysis

Preprocessing steps were performed using DataViewer (SR Research, Toronto, Ontario, Canada). Neighbor fixations less than 80ms were merged. Fixations less than 80ms or greater than 1000ms were excluded. Trials with blinks or track loss in the critical region were also discarded. For the interest areas, we identify and analyze data from three regions: the verb embedded in the relative clause (e.g., hoped), the adjective region (e.g., to be famous[3]) as the critical region where filled-gap effects are to be observed, and the prepositional phrase (e.g., by her parents) as a spillover and disambiguation region. We analyzed the following reading measures: first pass reading time, i.e., the sum of fixations from the first fixation into the region until the first exit out

[3] The infinitive region “to be” and the adjective region “famous” were treated as separated regions in the first place. Given the fact that the length of these two regions are relatively short and based on the observations that the reading time patterns were congruent with each other, we merged these two regions for data analysis.

of this region in either direction, representing the early stage of initial reading; regression path duration (go-past reading time),i.e., the sum of duration from the first fixation into the region until the exit out of the region to a later region, representing the later stage of initial readings, and total dwell time, i.e., the sum of all fixations fall in the region, representing the complete readings of sentences. These measures are typically known to be sensitive to the processing difficulties associated with filled-gap effects. These reading time measures were analyzed in linear mixed-effects models using the lme4 package in R (Bates et al., 2015). Pairwise comparisons were done using the emmeans package (Searle et al., 1980). Island type was included as a fixed effect, and subject, items were included as random effects. Maximal models were first applied and reduced effect slopes from random effects if the models failed to converge. Contrast code was created to make separate comparisons against no-island condition: negative-island condition vs. no-island condition; wh-island condition vs. no-island condition.

Results

Means and standard errors for the eye-movement measures in each region are reported in Table 3, and model summaries are reported in Table 4. Figure 3 demonstrates the results of the three reading times measures at the critical adjective region.

Verb region. Shorter reading time was observed in wh-islands compared with no-islands in first-pass reading time (Est. = 33.71, t = 3.10, p = 0.006) and total reading time

(Est. = 146.1, t = 2.49, p = 0.035). No significant effect was found in the comparison between negative-islands and no-islands.

Adjective region. Shorter reading times were found in wh-islands compared with no-islands in first-pass reading (Est. = 80.06, t = 2.60, p = 0.038), go-past reading (Est. = 145.9, t = 2.57, p = 0.028) and total reading (Est. = 242, t = 3.25, p = 0.009). However, no facilitations in reading time or regression out were found in negative islands compared with no-islands, in either first-pass reading (Est. = -7.45, t = -0.25, p = 0.967), go-past reading (Est. = 13.6, t = 0.24, p = 0.969) or total reading (Est. = 115, t = 0.918, p = 0.633).

Spillover region. Reading facilitations continued to the spillover region in wh-island conditions compared with no-island condition, reflected by shorter go-past reading time (Est. = 279.6, t = 2.31, p = 0.041) and total reading time (Est. = 179.96, t = 2.14, p = 0.05). For the comparison between negative-islands and no-islands, a marginal effect was found in first-pass reading time (Est. = -58.2, t = -2.04, p = 0.074) and in total reading time (Est. = -202.74, t = -1.73, p = 0.093) in the direction that negative islands require more reading time than no-islands.

	Verb region	Adjective region	Spillover region
First pass reading			
No-island	284(13)	480(24)	386(15)
Negative-island	270(9)	471(21)	448(19)
Wh-island	239(9)	394(17)	412(19)
Go-past reading			
No-island	386(52)	695(50)	977(82)
Negative-island	354(35)	716(45)	977(117)
Wh-island	428(37)	543(33)	701(69)
Total reading			
No-island	651(39)	1045(59)	802(45)
Negative-island	698(75)	1159(86)	912(71)
Wh-island	506(32)	802(45)	724(47)

Table 3. Means and standard errors for each measure in each region

	Eye-movement measures								
	First pass reading time			Go-past reading time			Total reading time		
	Est.	SE	t	Est.	SE	t	Est.	SE	t
Verb									
No vs. wh-island	33.71	10.9	**3.10***	-45.6	61.5	-0.74	146.1	58.8	**2.49***
No vs. neg-island	-7.79	10.8	-0.72	27.7	59.3	0.47	-45.6	58.6	-0.77
Adjective									
No vs. wh-island	80.06	30.8	**2.60***	145.9	56.8	**2.57***	242	74.4	**3.25***
No vs. neg-island	-7.45	30.5	-0.25	13.6	56.6	0.24	-115	125.6	-0.92
Spillover									
No vs. wh-island	-21.9	29.6	-0.74	279.5	121	**2.31***	180.0	84.0	**2.14***
No vs. neg-island	-58.2	28.6	-2.04	-3.18	121	-0.03	-202.7	117.0	-1.73

Table 4. Model summaries for each measure in each region

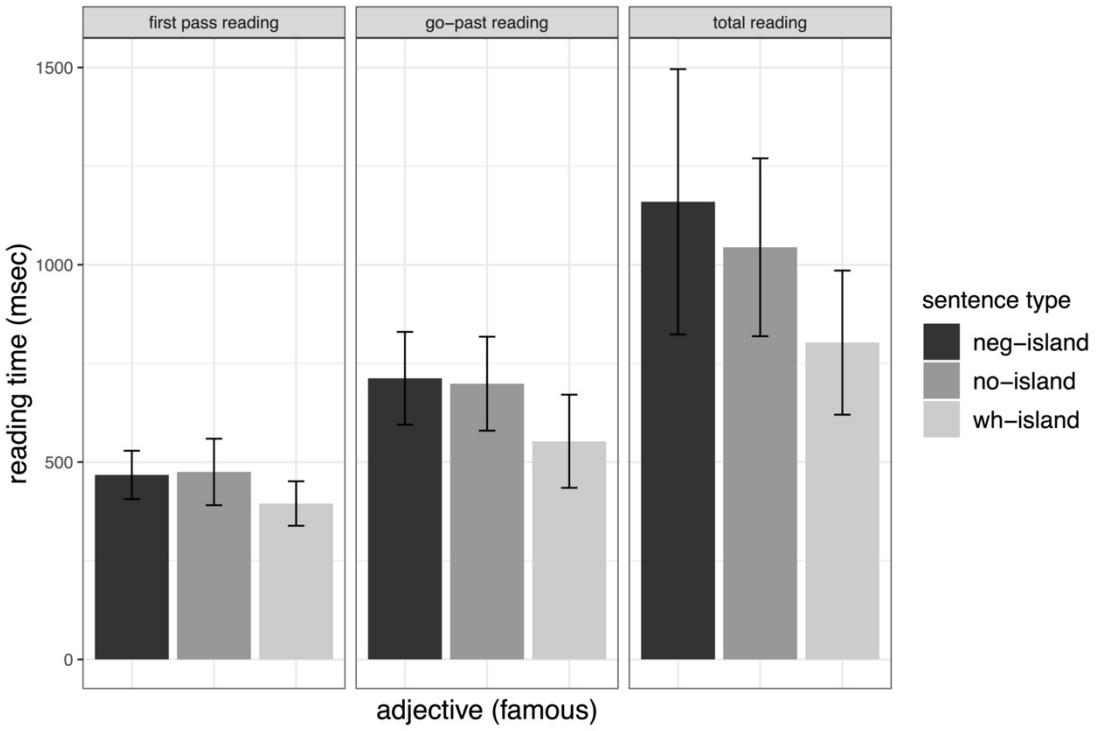


Figure 3. Mean reading time in each measure at the critical adjective region.

Discussion

To start with, in the comparison between wh-islands (syntactic islands) and no-islands, speedups in reading time in wh-islands were observed in all three measures (first pass reading time, go-past reading time and total reading time) at the critical adjective region where filled-gap effects typically show up when active gap filling is initiated. The speedups continued in the later spillover region in go-past reading time and total reading time. We interpret these findings as modulations of active dependency formation by strong islands. When the parser processes the previous relativizer region (i.e., “who was”) and recognizes the presence of an embedded relative clause island, it stops trying to associate the degree question filler (e.g., “how tall”) within the island domain, thus a filled adjective does not trigger filled-gap effects in wh-island conditions

as no anticipation of a gap is generated. This observation is consistent with previous studies claiming the parser is sensitive to strong island constraints. Additionally, reading facilitation in wh-islands also showed up at the verb region. One sensible explanation is that, the presence of a relativizer in wh-islands helps the parser to disambiguate the sentence structure compared with no-islands and negative-islands where a reduced relative clause creates a temporary ambiguity between relative clause structure and simple main clause structure. Thus, when processing the verb in wh-islands, the verb is unambiguously an embedded verb, and less processing time is required. When processing the verb region in no-islands and negative-island, the verb is temporarily ambiguous between a past-tense main verb and an embedded verb in a reduced relative clause structure.

What's more to our interest is the comparison between negative islands and no-islands. No significant effect was observed in negative island conditions in all three reading time measures at the critical adjective region, suggesting an agreeing pattern with no-island conditions. This alignment can be viewed as an indication that the parser does not recognize the negation as a cue for an island domain, thus treats negative islands as no-islands in real-time sentence processing and actively posits a gap, then filled-gap effects show up when an adjective takes up the hypothesized gap position. Besides the critical adjective region, the reading patterns of negative islands also demonstrate consistency with the reading patterns of no-islands in the spillover

/disambiguation region, suggesting that the comprehenders were experiencing the same temporal ambiguity of sentence structure and disambiguation as in no-islands. Some marginal effects of even longer reading times showed up in negative islands in first-pass reading and total reading potentially indicate that not only the presence of negation fails to help the parser to recognize an island domain, but also it, as an independent processing factor (see Gieselman et al., 2013), creates extra demands of resources from the parser to process the sentence structure, therefore, even longer reading times are expected for negative islands.

General Discussion

This study focuses on the real-time processing of negative islands, a type of weak island that selectively constrain wh-dependencies. Unlike strong islands which are syntactic in nature, negative island is in essence a presuppositional constraint of maximal informativity, i.e., the answer set contains a true answer entailing all the other true ones. While the processing of strong islands has long been well-studies, and the literature has established a point that human sentence processing parser is sensitive to strong island constraint, however, not much attention has been given to the parser's sensitivity to negative islands and it remains unclear whether island constraints arise from different levels will behave differently in affecting parsing decisions. The semantic/pragmatic level of calculations are considered to take more time and require more resources for the parser during the processing of the presupposition violation in

negative islands, but it has not been directly compared with strong islands in real-time measures. Novelly, by observing the effectiveness in blocking the filled-gap effect, we are able to directly compare different types of island constraints at the same time, to see the differences in time course when processing various island situations.

To start with, we confirmed that comprehenders are aware of negative island constraints offline in Experiment 1. The results show that, during off-line complete reading comprehension, negative degree questions are rated much lower than positive degree questions, suggesting participants realized the violation of maximal informativity and thus considered negative degree questions to be unacceptable. An interaction between polarity and sentence structure suggested that relative clause island effectively blocked the scope of negation, rescuing the acceptability of negation with relative clauses. More importantly, using real-time experimental measures, Experiment 2 and Experiment 3 demonstrate the differences in time course when parsing different types of islands. We first confirmed the typical findings that strong islands can be utilized by the parser to block the active gap searching within island domains in both experiments. Then we demonstrate that the negative islands, on the other hand, cannot block the filled-fap effect caused by active gap filling and the reading patterns of negative islands are consistent with no-islands. Experiment 2 found a filled-gap effect in wh-island at the critical region, the first spillover region, and the second spillover region. While no effect was found in negative islands at the critical adjective region or any spillover region,

though seemingly there was a trend of faster recovery from the filled-gap effect at the spillover region. Using more fine-grained reading time measures, Experiment 3 replicates and solidifies the findings about the strong island's blocking effect in Experiment 2 and also further clarifies that the trend of numerically faster recovery from filled-gap effect at the spillover region in negative island condition was spurious. It is evident by shorter reading times in strong island condition observed at the critical adjective region and spillover region with multiple reading time measures including first-pass reading time, go-past reading time, and total reading time, and by no significant difference between negative islands and no-islands in any region in all reading measures.

Our results demonstrate the differences in real-time processing between negative islands and wh-islands. These differences might be caused by different sources of island constraints. Wh-islands are more of syntactic constraints with clear syntactic boundaries that help comprehenders recognize the embedded domain, while negative islands arise from more semantic/pragmatic considerations which is not explicitly cued at the grammatical level. A grammatically explicit cue like a wh-phrase filler helps the parser recognize an embedded island domain whenever it presents itself in a sentence context. On the other hand, either a negator alone or a degree question alone is not an unambiguous cue for negative island constraints. A violation of maximal informativity context needs to be created to enable negative island constraints. Therefore, the exact degree question filler has to be fully retrieved upon processing a negator and

comprehenders need more time to calculate such a presupposition for questions and thus are unable to immediately recognize the cues for negative island constraints in real-time. Underlyingly this asymmetry between wh-islands and negative islands in real-time status might imply a dual-analyzer system for sentence processing that a heuristic parser cannot rapidly recognize a presuppositional constraint and such a violation can only be detected when boarder grammar knowledge comes into play later.

There are some studies going after a processing approach to understanding the nature of negative island phenomena, that negative island constraints are the overload of general processing costs. The central claim of processing-based theories (e.g. Deane, 1991; Hofmeister & Sag, 2010; Kluender, 1991; Kluender, 1998; Kluender, 2004; Kluender & Kutas, 1993) is that the avoidance of wh-extractions out of certain domains might be a result of complex configurations that exceed the capacity of working memory system, rather than the violation of grammatical constraints. The processing costs are more from an extra-linguistic perspective, that the unacceptability of negative questions is caused by individual factors that create accumulated processing difficulties, rather than by specific instantiation of grammatical knowledge. Gieselman and colleagues (2011) identified three factors that are known to increase processing difficulties in general: extraction, negation, and referentiality, and isolated each factor to see if these factors independently elicit differences in sentence acceptability with and without island contexts. In a series of acceptability tasks, they found robust effects of extraction types

(subject-extractions vs. object-extractions) and of referent types (which-phrase vs. how many-phrase) only in the presence of negation, while the effect of negation is prominent even in the absence of these other factors. They argue that negation is a general processing factor that creates difficulty in parsing complex structures and thus affects the acceptability of negative islands. However, it does not mean that island constraints can be reduced to processing factors that intervene in working memory load. There is also an increasing body of empirical studies suggesting that the island effects remain significant even when general processing factors are controlled in the experimental designs (e.g., Sprouse, 2007; Sprouse, Wagers, & Phillips, 2012). What's more, their data still come from offline acceptability tasks in which time-sensitive processing information is absent.

There also has been much debate over the strong/weak islands distinction, to summarize, about whether weak islands should be analyzed at the same level of representation as strong islands and whether weak islands and strong islands can be explained under the same mechanism. Underlyingly, our results indicate that negative islands might be at a different level of representation from strong islands and should be explained by separate mechanisms. However, one thing to note is that, although a negative island is generally categorized as a type of weak island, here we view the observed differences between wh-islands and negative islands as a result of the distinction of syntactic cues and semantic/pragmatic cues rather than strong/weak

distinction. Traditionally, the distinction between strong islands and weak islands is drawn based on the acceptance of various extraction types. Islands that prohibit all extractions are categorized as strong islands while those that cannot block all extractions are weak islands. Thus the group of weak islands is an unmarked category including various types of constraints, such as negative islands, extraposition islands, whether islands, etc. Some of them have a semantic/pragmatic nature while other weak islands are still syntactic islands. Therefore, regarding the discussion about whether there is a categorial difference between strong islands and weak islands, we think from a real-time processing perspective, the weak island type should not be treated as a unified group and the performance varies as different types of weak islands arise from different levels of representations. Our results suggest that negative islands and strong islands behave differently during real-time sentence processing due to their semantic/pragmatic properties. There are other studies suggesting that whether-island, as a type of weak island with a syntactic nature, is more in line with strong islands and demonstrates properties that can be rapidly utilized by real-time sentence parser.

Conclusion

In this paper, one off-line acceptability judgment and two real-time reading comprehension studies, one online self-paced reading task and one lab-based eye-tracking while reading task were conducted to investigate the human sentence parser's sensitivity to different types of island constraints. We first confirmed the comprehenders' awareness of negative island constraint using off-line judgment tasks. Then in the self-paced reading and eye-tracking studies, we replicated previous findings that the real-time parser is sensitive to wh-islands (syntactic islands), indicated by the blocking of filled-gap effects during active dependency formations. Novelly, we included negative islands in the comparison of the parser's sensitivity to island constraints, and provided evidence that negative island constraints cannot be rapidly used in real-time processing, indicated by the failure to block filled-gap effects during active dependency formations. These findings suggest that negative (semantic/pragmatic) islands might take time to emerge, unlike wh- (syntactic) islands which might take more immediate effects in real-time processing.

References

Abrusán, M. (2007). Contradiction and grammar: The case of weak islands: Massachusetts Institute of Technology dissertation.

Abrusán, M., & Spector, B. (2011). A semantics for degree questions based on intervals: Negative islands and their obviation. Journal of Semantics, 28(1), 107-147. doi: 10.1093/jos/ffq013.

Aoshima, S., Yoshida, M., & Phillips, C. (2009). Incremental processing of coreference and binding in Japanese. *Syntax*, *12*, 93–134.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphora. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 748–769.

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

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

Bock, J. K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*, 99–127.

Chomsky, N., Anderson, S., & Kiparsky, P. (1973). Conditions on transformations. In S. R. Anderson & P. Kiparsky (Eds.), A Festschrift for Morris Halle (pp. 232-286). New York: Holt, Rinehart, and Winston.

Chomsky, N. (1986). Barriers. MIT Press.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407.

Clackson, K., Felser, C., & Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: Evidence from eye-movements during listening. *Journal of Memory and Language*, *65*, 128–144.

Clifton C Jr and Frazier L (1989) Comprehending sentences with long-distance dependencies. In Carlson G and Tanenhaus M (eds) Linguistic structure in language processing. Dordrecht: Kluwer, pp. 94–128.

Clifton, C, Jr., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivista di* Linguistica, 11, 11–39.

Çokal, D., & Sturt, P. (2022). The real-time status of strong and weak islands. PLoS ONE, 17(2).

Comorovski, I. (1989). Discourse-linking and the wh-island constraint. North East Linguistics Society, 19(1), 7.

Cowart, W., & Cairns, H. S. (1987). Evidence for an anaphoric mechanism within syntactic processing: Some reference relations defy semantic and pragmatic constraints. *Memory and Cognition*, *15*, 318–331.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In A. M. Zwicky, D. R. Dowty, & L. Karttunen (Eds.), Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives (pp. 94-128). Cambridge University Press.

Dayal, V. (1996). Locality in wh-quantification. Dordrecht: Kluwer.

Deane, P. (1991). Limits to attention: A cognitive theory of island phenomena. Cognitive Linguistics, 2(1), 1-45.

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

Drummond, A. (2013). Ibex farm (free hosting for Internet-based experiments).

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, *8*(11), e81461.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. Current Directions in Psychological Science, 11(1), 11-15.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. Language and Linguistics Compass, 1, 71-83. doi: 10.1111/j.1749-818X.2007.00007.x

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. Linguistic Inquiry, 9(3), 427-473.

Fox, D., & Hackl, M. (2006). The universal density of measurement. Linguistics and Philosophy, 29(5), 537-586. doi: 10.1007/s10988-006-9004-4.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), Presupposition and implicature in compositional semantics (pp. 71-120). Palgrave MacMillan.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. Natural Language and Linguistic Theory, 5(4), 519-559.

Fujita H and Cunnings I (2020) Reanalysis and lingering misinterpretation of linguistic dependen- cies in native and non-native sentence comprehension. *Journal of Memory and Language* 115: 104154.

Gieselman, S., Kluender, R., & Caponigro, I. (2011). Pragmatic processing factors in negative island contexts. WECOL 2010, 59.

Gieselman, S., Kluender, R., Caponigro, I., Fainleib, Y., LaCara, N., & Park, Y. (2013).

Isolating processing factors in negative island contexts. In Proceedings of NELS (Vol. 41, pp. 233-246).

Gimenes, M., Rigalleau, F., & Gaonach, D. (2009). When a missing verb makes a French sentence more acceptable. Language and Cognitive Processes, 24, 440–449.

Grant, M., Sloggett, S., & Dillon, B. (2020). Processing ambiguities in attachment and pronominal reference. *Glossa: a journal of general linguistics*, *5*(1).

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. Language, 86(2), 366-415.

Huang, J. (1982). Logical Relations in Chinese and the Theory of Grammar. MIT dissertation.

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. The Quarterly Journal of Experimental Psychology, 69, 1013-1040. doi: 10.1080/17470218.2015.1053951

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backward anaphora. *Journal of Memory and Language*, *56*, 384–409.

Kluender, R. (1991). Cognitive constraints on variables in syntax. University of California, San Diego.

Kluender, R. (1998). On the distinction between strong and weak islands: a processing perspective. Syntax and Semantics, 29, 241-279.

Kluender, R. (2004). Are subject islands subject to a processing account. In Proceedings of WCCFL (Vol. 23, pp. 475-499). Somerville, MA: Cascadilla Press.

Kluender, R., & Gieselman, S. (2013). What's negative about negative islands? A re-evaluation of extraction from weak island contexts. Experimental syntax and island effects, 186-207.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbound dependencies. Journal of Cognitive Neuroscience, 5.

Kroch, A. (1989). Amount quantification, referentiality, and long wh-movement. Unpublished manuscript, University of Pennsylvania, Philadelphia.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. Journal of Psycholinguistic Research, 44, 27-46.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehen- sion. *Journal of Psycholinguistic Research*, *18*, 5–19.

Omaki A and Schulz B (2011) Filler-gap dependencies and island constraints in second language sentence processing. Studies in Second Language Acquisition 33: 563–88.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*, 427–456.

Phillips, C. (1996). Order and structure (Doctoral dissertation, Massachusetts Institute of Technology).

Phillips C (2006) The real-time status of island phenomena. *Language* 82: 795–823.

Pickering, M., Barton, S., & Shillcock, R. (1994). Unbounded dependencies, island constraints and processing complexity. *Perspectives on sentence processing*, 199-224

Rizzi, L. (1990). Relativized minimality. Cambridge, MA: MIT Press.

Rizzi, L. (2003). Relativized minimality effects. In M. Baltin & C. Collins (Eds.), The handbook of contemporary syntactic theory (pp. 89-110). Oxford, UK: Blackwell.

Rizzi, L. (2004). Locality and left periphery. In A. Belletti (Ed.), Structures and beyond: The cartography of syntactic structures (Vol. 3, pp. 223-251). Oxford, UK: Oxford University Press.

Ross, J. R. (1984). Inner islands. Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society, 258-265.

Ross, J. R. (1967). Constraints on variables in syntax (Doctoral dissertation, Massachusetts Institute of Technology).

Rullmann, H. (1995). Maximality in the Semantics of Wh-constructions. University of Massachusetts Amherst.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. The American Statistician, 34, 216-221.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. Biolinguistics, 1, 123-134.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. Behavior research methods, 43, 155-167.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. Language, 88, 82-123.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. Language and Cognitive Processes, 1, 227-245.

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

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*, 201-216.

Szabolcsi, A., & Zwarts, F. (1993). Weak islands and an algebraic semantics for scope taking. Natural Language Semantics, 1, 235-284.

Townsend, D. J., & Bever, T. G. (2001). Sentence comprehension: The integration of habits and rules. Cambridge, MA: MIT Press.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. Journal of Memory and Language, 35(3), 454-475.

Trotzke, A., Bader, M., & Frazier, L. (2013). Third factors and the performance interface in language design. Biolinguistics, 7, 1–34.

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

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

Yoshida, M., Aoshima, S., & Phillips, C. (2004). Relative clause prediction in Japanese. In *Talk at the 17th* annual CUNY conference on human sentence processing. College Park, MD.