Evidence from behavioral experiments:

Information theory and discourse-based accounts of long-distance dependencies

A dissertation presented

by

YINGTONG LIU

to

THE DEPARTMENT OF LINGUISTICS

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in the subject of

LINGUISTICS

Harvard University

Cambridge, Massachusetts

March 2022

Dissertation Advisors: Edward Gibson & Jesse Snedeker Yingtong Liu

Evidence from behavioral experiments:

Information theory and discourse-based accounts of long-distance dependencies

Abstract

For decades, linguists and psychologists have sought to understand why some long-distance dependencies sound grammatical while others less so, and how people process them. In this dissertation, I investigate these puzzles and get promising results, using a variety of behavioral experiments and statistical modeling. This dissertation consists of three experimental projects, examining how various factors from language exposure, communicative pressure, discourse, syntax and semantics shape people's acceptability and interpretation of long-distance dependencies. I find that our frequency-based processing proposal provides a more succinct explanation for the observed acceptability data of the tested island phenomena than the previous discourse, syntactic and semantic accounts. In addition, I show that experimental data reflecting the referential properties of Chinese anaphors support the discourse-based logophoricity accounts, not the pure syntactic theories. I have also better characterized the structural prior and the source of non-literal interpretations in noisy-channel processing of filler-gap constructions.

Table of Contents

Acknowledgements

I feel grateful to many people who have made this dissertation happen.

First and foremost, I would like to thank my dissertation committee members.

To **Ted**: Thank you for introducing the world of psycholinguistics to me and guiding me through the past six years. You have been an incredible advisor to me – always encouraging, insightful, humorous, and generous with your time and resources. I was completely at lost at grad school when I walked into your class. Over the years, under your advising, I've grown into an independent experimenter and researcher. I can hardly express my gratitude to you with words. Your influence is on every page of this dissertation, and beyond. Thank you and Ev for all the Thanksgiving dinners and fun games, which has made this foreign land homelike to me.

To **Jesse**: I feel incredibly grateful for your taking a chance on me and being my committee chair. You were the first faculty I talked to in grad school, after my twelve-hour flight back to 2016, encouraging me to work on sentence processing. Looking back, all the dots are connected with you. You are always dedicated, patiently listening to my ideas, and challenging me intellectually. Paper 3 couldn't be in its current great shape without your suggestions. Thank you for the numerous hours you devoted to me and my projects, for making me a better researcher, and for supporting me exploring various career paths.

To **Kevin**: I feel extremely lucky to have you on my committee. Your trust has supported me walking through the difficult times. You always give me the freedom to explore areas that I'm interested in, and believe that I can do it, even if that's far away from canonical P-side studies. Thank you for dedicating so much time and patience to learning about me and my work. I couldn't have made this journey without your trust.

Alongside my committee members, I would like to thank Isabelle Charnavel. Your advising, support and help have shaped Paper 2 in this dissertation. Your insights from the theoretical side and open-mindedness towards experimental approach are always inspiring.

I would like to thank my amazing collaborators, Rachel Ryskin and Richard Futrell. I learned so much from you two, including but not limited to experimental design, data analysis, abstract and paper writing, and how to be a researcher. Paper 1 and Paper 3 in this dissertation couldn't have been possible without you.

I want to thank my support networks, especially: Giuseppe Ricciardi, Shannon Bryant, Zachary Rothstein-Dowden, Yujing Huang, Fangdai Chen, Tianwang Liu, Yuyin He, Jayden Ziegler, Sherry Chen, Yenan Sun, Lina Nie, Weirong Guo, Simge Topaloglu, Cristina Aggazzotti, Alex Paunov, Saima Malik-Moraleda, Dora Mihoc, Gregory Scontras, Josh Martin, Gasper Begus, Yimei Xiang, Lena Boris, Aurore Gonzalez, Zuzanna Fuchs, Dorothy Ahn, Yuhan Zhang, Zheng Zhang, Gunnar Lund, Pooja Paul, Yingzhao Zhou, Yuhang Xu, Ian Kirby, Ethan Wilcox, Niels Kuehlert, Jack Rabinovitch, Tiffany Yang, Ankana Saha, Deniz Satik, Tamisha L. Tan, Yige Wang, Ran Wei, Kexin Yi, Jun Yin, Hengyun Zhou, Peng Qian, Anna Ivanova, Hope Kean, Anthony Yacovone, Maggie Kandel, Joe Coffey, Dejing Liu, Yipu Wei, Lee Ling Ting, Hai Hu, and the rest of TEvLab and SnedLab families. Your friendship and accompany have made this journey much more delightful.

Thanks to my previous advisors at Renmin University of China and Utrecht University for encouraging me to pursue the doctoral study. I owe great gratitude especially to Qingmin Guo, Martin Everaert, Eric Reuland, Joost Zwarts, Craig Thiersch, and Marijana Marelj.

Lastly, **Mom and Dad**, thank you for your unreserved love and support to me for all the paths I've chosen, even if that means living far away from you for years. If I have accomplished

something, the triumph is yours. I love you, and I can't wait to hug you in person. My greatest gratitude is reserved for **Chi**. I feel incredibly fortunate to have met you. It all started in a sunny summer afternoon, right in front of Perkins, and then you joined me on my journey. Thank you for my favorite letters, the yummiest dishes, and our wild explorations of nature, for always encouraging the right things in me, and for driving me everywhere. You lighted the darkness in my most trying times.

This dissertation is dedicated to my parents and my husband Chi.

Chapter 1: Introduction

The main body of this dissertation comprises three papers, investigating two types of long-distance dependencies – filler-gap construction and anaphora - from the perspectives of processing, discourse, syntax, and semantics. Below, I summarize the three studies presented in this dissertation.

1.1 Paper 1

Going back to Ross (1967) and Chomsky (1973), researchers have sought to understand what conditions permit long-distance dependencies in language, such as between the wh-word *what* and the verb *bought* in the sentence 'What did John think that Mary bought?'. In Paper 1, we attempt to understand why changing the main verb in wh-questions affects the acceptability of long-distance dependencies out of embedded clauses. In particular, it has been claimed that factive and manner-of-speaking verbs block such dependencies (e.g., 'What did John *know/whisper* that Mary bought?'), whereas verbs like *think* and *believe* allow them. We provide 4 acceptability judgment experiments of filler-gap constructions across embedded clauses to evaluate four types of accounts based on (1) discourse; (2) syntax; (3) semantics; and (4) our proposal related to verb-frame frequency. The patterns of acceptability are most simply explained by two factors: verb-frame frequency, such that dependencies with verbs that rarely take embedded clauses are less acceptable; and construction type, such that wh-questions and clefts are less acceptable than declaratives. We conclude that the low acceptability of filler-gap constructions formed by certain sentence complement verbs is due to infrequent linguistic exposure.

1.2 Paper 2

Anaphors like Icelandic *sig* and Mandarin *ziji* do not obey Binding Condition A (Chomsky 1986). Two competing approaches have been proposed to capture binding observations that violate Condition A: (i) The long-distance binding (LDB) theory derives non-local binding via covert cyclic movement and turns non-local binding into local binding which obeys Condition A (Pica 1987, Huang & Tang 1991); (ii) The logophoricity theory explains exempt anaphors by *logophoric* rather than structural constraints (Sells 1987, Huang & Liu 2001, Charnavel 2019). The two theories make distinct predictions about referential dependencies between reflexives and their antecedents. The LDB theory predicts that antecedents should always c-command the reflexives, while the logophoricity theory predicts that reflexives need not be c-commanded by their antecedents if they are logophoric. Paper 2 tests the two competing theories by investigating the referential properties of Chinese reflexives *ziji* and *taziji* using three acceptability judgment tasks. The results regarding *ziji* support the logophoricity theory over the LDB theory. Furthermore, the results reveal that *taziji*, though usually considered as a local anaphor (e.g., Cole, Hermon & Huang, 2006), can be exempt from binding under logophoric conditions, like *ziji*.

1.3 Paper 3

Comprehenders must recover a speaker's intended meaning from linguistic input that may be corrupted by noise. Recent work suggests that they do so by combining prior knowledge $P(S_i)$ about what sentences are likely to be intended, with a noise model $P(S_i \rightarrow S_p)$ which encodes how an intended sentence S_i might be corrupted to a perceived sentence S_p in transmission, as in (1)

(Gibson et al., 2013; Poppels & Levy, 2016; Ryskin et al., 2018). For example, although *the cheese* should be the agent based on the form in 'It was the mouse that the cheese ate.', around 42% of comprehenders might infer that an error occurred between the speaker's intent and the ultimately-perceived sentence, such that *the mouse w*as the actual agent.

(1) $P(s_i|s_p) \propto P_L(s_i)P_N(s_i \rightarrow s_p)$

Previous studies have tested the noisy-channel framework by manipulating the prior probability of the *meaning* of the intended sentence (i.e., plausibility), but the theory also predicts that the *form* of the sentence matters. We thus re-define the language prior $P_L(s_i)$ as the joint probability of the meaning prior $P_L(s_{i\_meaning})$ and the structure prior $P_L(s_{i\_grammar})$, as in (2) and (3). According to our proposed noisy-channel model (3), both meaning plausibility and structural frequency affect comprehenders' interpretation of the input sentence. Sentences with low structural prior can trigger more non-literal interpretations, as the neighborhood candidates with higher structural prior might be considered as the actual intended utterance.

(2) $P_L(s_i) = P_L(s_{i\_grammar}, s_{i\_meaning})$

(3) $P(s_i|s_p) \propto P_L(s_{i\_grammar}, s_{i\_meaning})\, P_N(s_i \rightarrow s_p)$

Paper 3 tests how English and Mandarin speakers interpret simple transitives and clefts of plausible and implausible meanings, with the goal of investigating three research questions: (i) Are comprehenders more likely to draw inferences for sentences formed in low frequency structure than those formed in high frequency structures, as predicted by our proposed model in

(3)? (ii) What 'grain sizes' distributional syntactic information is stored by language users? And (iii) where do the non-literal interpretations come from?

I find that (a) there are significantly more non-literal interpretations for low-frequency structures than for high-frequency structures in both simple transitives and clefts, confirming the impact of structural frequency on noisy-channel processing; (b) the amount of non-literal responses is in proportion to the relative frequency contrast between the input and its neighborhood candidates, suggesting comprehenders track the construction, rather than the linear string of content words in the input sentence; (c) comprehenders can correctly retype the input sentence among over 99% responses, including trials with non-literal responses. That indicates that comprehenders' non-literal responses do not come from their misperception of the input sentence, and the non-literal interpretations are more likely to come from the comprehenders' inference about the speaker's intended utterance, which aligns our noisy-channel proposal. All these observations are replicable in our experiments. Thus, these results provide exceptionally robust evidence for our proposed noisy-channel model with integration of syntactic information.

Chapter 2: [Paper 1] A verb-frame frequency account of constraints on long-distance dependencies in English

2.1 Introduction

An important feature of human languages is that they contain constructions that license long-distance dependencies: so-called *filler-gap* constructions, such as wh-questions, relative clauses, clefts and topicalization in English and other Germanic languages, and in many other language families. These constructions involve a displaced constituent -- a *filler* -- that appears in a position other than its canonical position in a declarative clause. The place where the constituent would appear in a declarative is known as the *gap* site, which we will indicate with an underscore ‘‘_’’. For example, the declarative form of a simple clause is provided in (1a), along with a wh-question version of this clause in (1b), where the patient (object) is fronted. A corresponding relative clause is provided in (1c) and a cleft is in (1d)[1]:

(1) a. John said that Mary bought the apple.

b. wh-question: What$_i$ did John say that Mary bought __$_i$?

c. relative clause: The apple that$_i$ John said that Mary bought __$_i$

d. cleft: It was the apple that$_i$ John said that Mary bought __$_i$

While the long-distance dependencies in (1) are possible, others are less acceptable, as in (2) (Ross, 1967; Chomsky, 1973). In the theoretical literature, the less acceptable versions in (2) have been called ‘islands’ to extraction: unacceptable long-distance filler-gap constructions.

[1] Following standard notation in the linguistics literature, we will notate the position in the declarative that is associated with the fronted element with an empty element “__”. We provide a subscript such as “i” to the fronted element (the “filler”) and the empty position. This corresponds to what movement-based theories call a gap or trace (Ross, 1967; Chomsky, 1973) but we use it mainly for ease of exposition (see Sag, 2010, for a traceless analysis).

(2) a.* Who$_i$ did [$_S$ you hear [$_{NP}$ the statement that the CEO promoted __$_i$]] ?

b. * Who$_i$ do [$_S$ you think [$_{NP}$ the gift from __$_i$] prompted the rumor] ?

c. * The bread that$_i$ [$_S$ you heard [$_{NP}$ the statement that Jeff baked __$_i$]]

d. * The politician who$_i$ [$_S$ you think [$_{NP}$ the gift from __$_i$] prompted the rumor].

In experimental investigations of the acceptability of materials involving long-distance dependencies like these, many researchers have also evaluated control materials with shorter dependencies (3a, b), and materials without the potential intervening material (3a, c), relative to the "island" structure in (2a)/ (3d):

(3) a. short, simple: Who heard that the CEO promoted the manager?

b. short, complex: Who heard the statement that the CEO promoted the manager?

c. long, simple: Who did you hear that the CEO promoted?

d. long, complex (the "island" structure): Who did you hear the statement that the CEO promoted?

In Sprouse et al. (2012, 2016), it is shown that the extracted complex version in (3d) is rated much worse than the other 3 conditions (a-c), resulting in a super-additive interaction between the two factors (Figure 1).

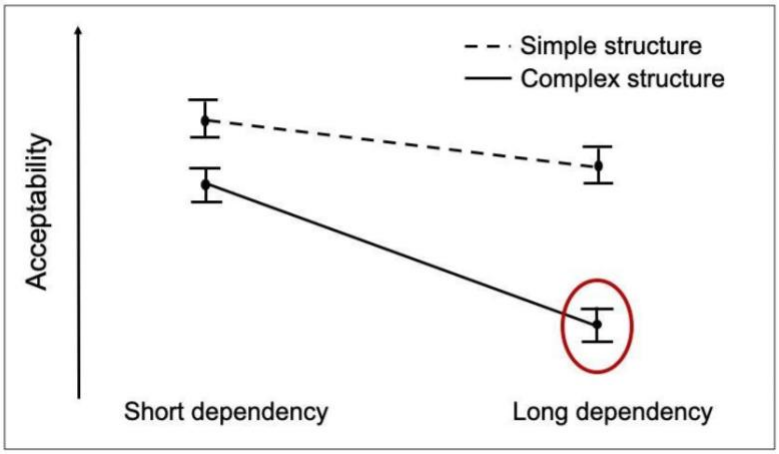


Figure 1. Illustration of a super-additive island effect, such that the complex, long dependency structure is rated least acceptable of the four conditions, and there is an interaction between dependency length (short vs. long) and complexity of the structures (complex vs. simple).

Several studies have followed Sprouse et al. (2016) in assuming that superadditivity as in Figure 1 effectively defines island-hood (e.g., Kush et al., 2019), with the consequence that an island is an unacceptable structure for which the source of unacceptability is not yet understood.[2] We will not make this assumption here, because this use of the term "island" presumes knowledge (or lack of knowledge) of the source of the unacceptability. For simplicity, we will therefore refer to unacceptable long-distance filler-gap constructions as islands, whether or not the reason for their unacceptability is known (Liu, Winckel et al., 2021).

The major theoretical interest in island phenomena began with Chomsky (1964, 1973), who argued that because extractions were similarly impossible across a range of constructions with different meanings (e.g., wh-questions, relative clauses, cleft structures, etc.), the constraints on extraction must be based on their syntactic form (see also Chomsky, 1977, 1981, 1986; Huang, 1982; Rizzi, 1990). Thus, Chomsky

[2] It is often assumed that some kind of syntactic constraint is responsible for the unacceptability, but so far no empirical independent evidence has been provided for such an assumption, largely because studies that sought to provide independent evidence for this assumption were mostly designed to filter out a subset of alternative explanations rather than directly testing the syntactic hypothesis (for details see Liu, Winckel et al., 2021).

argued for a pure structural account, which was called *Subjacency.* According to the details of that account, noun phrase (NP) and sentence (S) syntactic nodes are defined to be *bounding nodes* for extraction. Extraction across two bounding nodes was proposed to be ungrammatical. Consequently, the extractions in (2a-d) result in unacceptable sentences.

Furthermore, Chomsky argued that these constraints are unlearnable and hence innate, because of a classic poverty of the stimulus argument Chomsky (1973; 1981; 1986b): (a) extractions are unacceptable independent of the meaning of the constructions involved; and (b) a child would not be exposed to the right input across all the different constructions in which they hold - she is only exposed to examples of acceptable sentences, and there is no instruction with direct negative evidence (Hoekstra & Kooij 1988; Newmeyer 1991; see Ambridge, Pine & Lieven 2014 for a critical view).

In this paper we focus on extractions out of sentence complements of factive and manner-of-speaking verbs, as in (4). Researchers have long noted that extractions out of sentence complements taken by factive verbs – such as "know" (4b), "regret", and "notice", the contents of which are presupposed (Kiparsky and Kiparsky, 1971) – and sentence complements of manner-of-speaking verbs – such as "whisper" (4c) "mutter", and "mumble", which describe physical characteristics of the speech act (Zwicky, 1971) – are less acceptable than extractions across "bridge" verbs such as "say" (4a), "think" or "believe". Hence, the embedded clauses of factive and manner-of-speaking verbs are called 'islands', which are reported to ban extraction (e.g., Erteschik-Shir, 1973; Snyder, 1992; Ambridge & Goldberg, 2008; cf. individual differences in how good the baselines are; Dabrowska, 2010).

(4) a. Bridge verb

What did John **say** that Mary bought?

b. Factive verb

??What did John **know** that Mary bought?

c. Manner-of-speaking verb

??What did John **whisper** that Mary bought?

Note that what constitutes a "bridge" verb is not independently defined in the literature: a bridge verb is simply one for which extraction from its sentence complement is possible.

Below we review the three types of existing theories which aim to capture acceptability variance for extractions across various sentence complement verbs, and introduce our verb-frame frequency account.

2.1.1 Three types of existing theories and a new verb-frame frequency account

The three types of existing accounts are the information structure, syntactic, and semantic accounts.

A. Information Structure Accounts

Information structure refers to how information is packaged for the listener (e.g., Ambridge & Goldberg, 2008; Deane, 1991; Erteschik-Shir, 1973, 1979, 1998; Goldberg, 2006; Goldberg, 2016; Van Valin, 1998; Van Valin & LaPolla, 1997). Grammatical constructions specify certain parts of a sentence as 'focused' or 'backgrounded': Focused constituents are the main assertion of the sentence, while other parts of the sentence convey less salient information, and are therefore 'backgrounded'. According to this kind of proposal, wh-questions can't ask about backgrounded constituents, because that would lead to a clash of the function of wh-questions and backgrounded constructions: the wh-word is a classic focus, while constituents in backgrounded constructions cannot be focused. A constituent cannot felicitously be both discourse-prominent and backgrounded at the same time (Goldberg, 2016).

In this spirit, Ambridge & Goldberg (2008; henceforth A&G) proposed an account they call Backgrounded Constituents are Islands (BCI), as in (5). Extraction from a sentence complement is unacceptable in proportion to its backgroundedness: the more backgrounded the embedded clauses, the less acceptable the extraction.

(5) *Backgrounded Constituents are Islands (BCI)*:

Backgrounded constituents may not serve as gaps in filler-gap constructions.

In order to distinguish backgrounded constituents from focused constituents, A&G (2008) proposed the negation test. According to this test, the more backgrounded a constituent of a sentence is, the less likely that sentential negation can fall on it. Thus, a clause that is unlikely to be negated by sentential negation is more likely to be backgrounded, and is therefore more likely to ban extraction. Thus, factive verbs take the most backgrounded sentence complements, presuppositions, as in (6a), so the negation in the matrix clause does not affect the presupposed embedded clause. In contrast, the embedded clauses taken by bridge verbs are assertions and not backgrounded at all. For instance, in (6c), the sentential negation in Sentence 1 can negate the embedded clause. The backgroundedness of manner-of-speaking embedded clauses is claimed to be intermediate (6b).

(6) a. Sentence 1: I **didn't know** that Mary bought a car.

$\not\rightarrow$ Sentence 2: Mary didn't buy a car.

b. Sentence 1: I **didn't shout** that Mary bought a car.

?$\not\rightarrow$ Sentence 2: Mary didn't buy a car.

c. Sentence 1: I **didn't think** that Mary bought a car.

→ Sentence 2: Mary didn't buy a car.

Examples that support the BCI account include unacceptable extractions from a complex NP (7a) and sentential subject (7b). The relative clause 'who met ….' in (7a) is more backgrounded compared to the head noun 'the boy', and therefore bans extraction. Though the subject of a sentence is relatively salient in discourse – the default *topic* - constituents within a subject are also backgrounded as they are not themselves the primary *topic*.[3] Thus extraction out of a subject is not allowed as in (7b).

(7) a.*Who$_i$ did she see [the boy who met __$_i$]?

b. ??Who$_i$ did [that she hit __$_i$] was horrible?

(Examples from Goldberg, 2016)

A&G (2008) provided supportive evidence for the BCI account. They found a strong negative correlation between the negation test scores and difference rating scores between wh-questions and their corresponding declarative clauses ($r=-0.83$, $p= 0.001$). Factive verbs had the highest negation scores and difference scores, yielding the strongest islands. Bridge verbs had the lowest negation scores and difference scores, forming the weakest islands. Negation and difference scores for manner-of-speaking verbs were in the middle. However, these results only included a limited set of 12 verbs.

B. Syntactic Accounts

[3] Subject is the default *topic* of a clause, and what a sentence is 'about' (Chafe, 1987; Goldberg, 2016; Lambrecht, 1994; MacWhinney, 1977). That is, a clausal topic is a "matter of [already established] current interest which a statement is about and with respect to which a proposition is to be interpreted as relevant" (Michaelis & Francis, 2007). For extraction out of subject, see Abeillé et al. (2020) for a related but different perspective.

In order to explain the difference between extraction across bridge verbs on the one hand (4a) and extraction across factive and manner verbs on the other (4b/c), a syntactic account proposes different syntactic structures for bridge verbs compared to the other two kinds of verbs. It has been claimed that bridge verbs take embedded clauses as arguments, while embedded clauses of manner-of-speaking verbs and factive verbs contain extra covert structures at an abstract level ('Deep Structure' in Chomsky's framework) (cf. Baltin, 1982; de Cuba, 2018; Kiparsky & Kiparsky, 1971; Snyder, 1992; Stowell, 1981; Stoica, 2016). More specifically, Snyder (1992) argued that the underlying syntactic representation (8b) with manner-of-speaking verb *grunt* is actually (8a), and the clausal complement is covertly a modifier of the NP '*(a) grunt*'. Kiparsky & Kiparsky (1971) hypothesized that there is a covert *the fact* for factive verbs in the Deep Structure rendering the sentence complement part of a complex NP[4], as shown in (9). Assuming that complex NPs and adjuncts disallow extraction (Chomsky, 1981, 1986; Huang, 1982), (4b) and (c) could be ruled as ungrammatical under such a hypothesis.

(8) a. I **[$_{lightV}$(made)][$_{NP}$ (a) [$_{N}$ grunt]]**, (that is) Mary bought a car. (Deep Structure)

b. I **grunted** that Mary bought a car. (Surface Structure)

(9) a. I regret **the fact** that John bought a car. (Deep Structure)

b. I regret that John bought a car. (Surface Structure, via *fact*-deletion)

In this way, the unacceptability of extraction across factive and manner-of-speaking verbs could be captured by syntactic constraints of extraction such as *Subjacency*, which are hypothesized to be innate.

[4] One motivation for this proposal was that only factive verbs can overtly take 'the fact that…' (Kiparsky & Kiparsky, 1971), but some bridge verbs can also take this phrase (e.g., 'Mary *reported* the fact that France won the 2018 World Cup.').

But a serious problem with this kind of account is that there are no independent reasons to propose these covert complex structures.

C. Semantic Accounts

It has been proposed that sentence complement verbs may be categorized into two groups: factive and non-factive verbs. Sentence complements of factive verbs are presuppositions and non-factive verbs do not take presuppositions (e.g., Kiparsky & Kiparsky, 1971). A natural explanation for the acceptability contrast between bridge and factive wh-questions could be that presupposition does not allow extraction, while non-presupposition does.

There are three potential issues with this account. First, there has never been an independent basis for what counts as a 'bridge' verb, which calls into question meaning-based solutions to the puzzle of what makes such extractions possible. Second, the notion of *factivity* seems to be gradient rather than binary (Tonhauser, Beaver, & Degen, 2018), and therefore it is hard to find a clear boundary between 'factive' and 'non-factive' verbs. Third, manner-of-speaking verbs are not factive, so they should be grouped with bridge verbs, since neither of them take presuppositions[5]. Thus, this account may not be able to cover the contrast between extraction across bridge and manner-of-speaking verbs.

D. Our Verb-frame Frequency Account

We propose that the acceptability of filler-gap constructions involving extraction across sentence complement verbs and their corresponding declaratives can be explained by two independent, additive

[5] Kiparsky & Kiparsky (1971) didn't further sub-categorize the non-factive verbs. Given the provided threshold, bridge and manner-of-speaking verbs should both belong to the group of non-factive verbs.

factors, as in (10). One factor is the frequency or the type of the construction. Wh-questions are rated less acceptable than declaratives, because wh-questions are less common than declarative statements (Roland et al., 2007).[6] The second factor is the frequency of the verb head-structure: the joint probability of the verb *x* and *x* taking a sentence complement, in the form of P(matrix verb, sentence complement), as in (10).

(10) *The Verb-frame Frequency Hypothesis*:

The acceptability of a sentence is best captured by two independent effects: (i) the frequency or the type of the construction (e.g., wh-questions vs. declaratives) and (ii) the frequency of the verb head-structure, P(matrix verb, sentence complement) = P (matrix verb) * P (sentence complement | matrix verb).

This idea builds on Dabrowska (2008), who proposed that speakers store prototypical templates corresponding to frequent combinations such as 'Wh-word *do you think/say* sentence-complement?', such that filler-gap constructions that are more similar to prototypical constructions are more acceptable. A&G (2008) tested Dabrowska's proposal by means of a correlation analysis for wh-question acceptability and ratings of similarity of the main verbs involved to '*think'* or '*say'.* Their results showed no reliable correlation between semantic-similarity judgment data and well-formedness of wh-questions for either '*think'* (r=0.08, p=0.79) or '*say'* (r= 0.17, p=0.62), which casts doubt on the specific proposal of Dabrowska (2008).

Unlike Dabrowska's proposal, our proposal is not about any particular common verb. Rather, we build on previous work that has shown that less frequent or unpredictable extractions are more difficult to

[6] Other cognitive constraints, such as extra processing cost associated with filler-gap constructions, may also play a role (e.g., Hofmeister & Sag, 2010).

process (Kothari, 2008; Hale, 2001, 2003; Jurafsky, 2003; Levy, 2008; Verhagen, 2005), so that the unacceptability of certain filler-gap constructions might be due to infrequent exposure. Specifically, Kothari (2008) demonstrated that there is no categorical acceptability distinction between wh-questions formed by manner and non-manner of speaking verbs; instead, what matters more might be frequencies measured based on the verb, such as lemma frequency or subcategorization frequency.

According to our proposal, manner-of-speaking and factive wh-questions are less natural because the joint probability of those verbs and their taking sentence complements is lower. If they do take sentence complements with a similar frequency to bridge verbs, then they should form equally good wh-questions. In this way, within-verb group variance and across-verb group overlap in wh-question acceptability can be captured in this account.

2.1.2 Predictions of the four theories on factive and manner-of-speaking islands

The four accounts make distinct predictions about the acceptability patterns of filler-gap constructions formed by various sentence complement verbs.

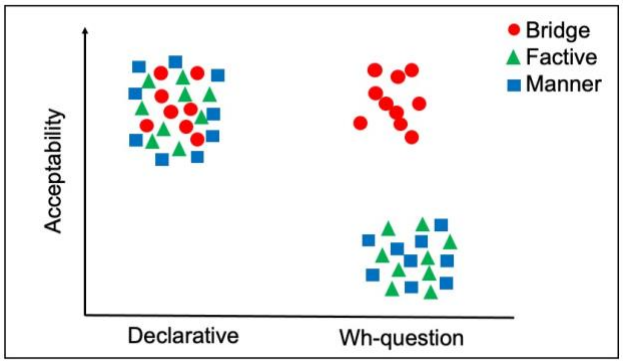
The syntactic accounts predict that all factive and manner-of-speaking wh-questions should be less acceptable than all the bridge ones due to categorically distinct covert structures which forbid extraction (e.g., Kiparsky & Kiparsky, 1971; Stowell, 1981; Snyder, 1992), as in Figure 2a.

The semantic accounts predict that all factive wh-questions are less acceptable than all the bridge and manner-of-speaking ones, as shown in Figure 2b, because only factive verbs take presuppositions, non-factive verbs do not. Extraction out of presuppositions should be less acceptable than out of non-presuppositions (Kiparsky & Kiparsky, 1971).

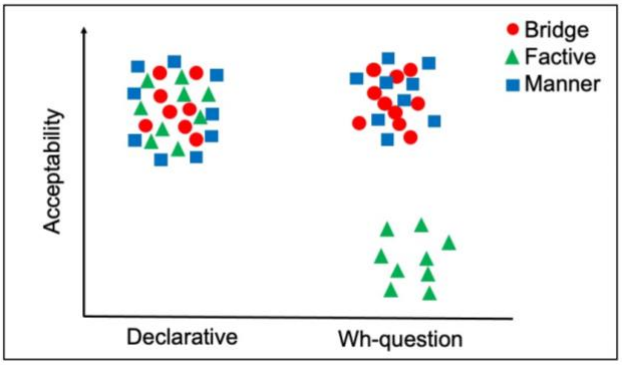
The BCI account (A&G 2008) predicts that the more backgrounded the sentence complement is, the less acceptable the wh-question. A&G (2008) measured wh-question acceptability by calculating the difference score between ratings of declaratives and the corresponding wh-questions -- higher difference scores indicate low acceptability -- and backgroundedness of the sentence complement using the negation test -- lower negation test scores suggest strong backgroundedness. Thus, following A&G (2008), there should be a strong negative correlation between difference scores and negations scores, as in Figure 2c. Factive verbs take presuppositions, the most backgrounded constituents, and therefore should receive the lowest negation scores and highest difference scores (lowest acceptability). Manner-of-speaking verbs should form more natural wh-questions, while bridge verbs construct fully acceptable wh-questions.

The verb-frame frequency account makes two predictions. First, the effect of verb-frame frequency should be similar for both declaratives and filler-gap constructions, resulting in no interaction. Second, within declaratives or filler-gap constructions, the higher the verb-frame frequency, the more acceptable the sentence, as plotted in Figure 2d.

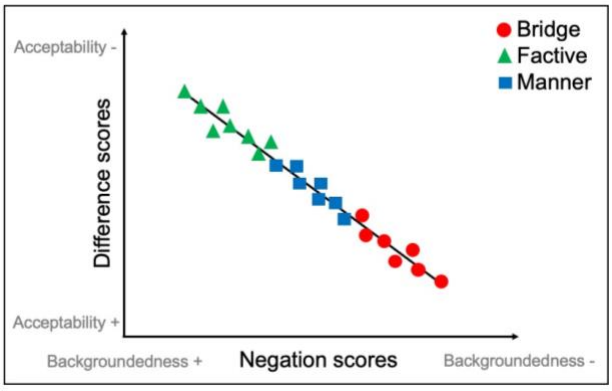
2a. Predictions of the syntactic accounts.



2b. Predictions of the semantic accounts.



2c. Predictions of the discourse BCI account.



2d. Predictions of the verb-frame frequency account.

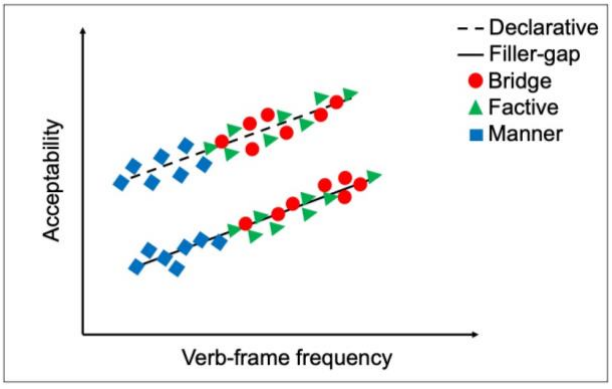


Figure 2. Predictions of the syntactic, semantic, discourse and frequency accounts. Each dot represents a word (conceptually). In Figures 2a, 2b and 2d, the y-axis is the raw rating. In Figure 2c, the y-axis denotes the difference scores between ratings of wh-question and declaratives (following Ambridge & Goldberg, 2008).

The remainder of this paper is structured as follows. Experiment 1 is a replication and extension of A&G (2008) in which we evaluated the existing discourse, syntactic, and semantic accounts. The predictions of these accounts are not consistent with our observed data. We therefore conducted post-hoc analyses of Experiment 1 to test our proposed verb-frame frequency account. Experiments 2 and 3 provide further support for the verb-frame frequency account with an extended set of sentence complement verbs and two filler-gap dependency constructions -- wh-questions and cleft structures.

2.2 Experiment 1: Replication of Ambridge and Goldberg (2008)

In Experiment 1, we attempted a replication and extension of A&G (2008) using an expanded set of 24 verbs in the 3 categories (A&G tested 12 verbs). There were two sub-experiments: (a) Experiment 1a which consisted of acceptability judgements of wh-questions formed by the 3 groups of verbs and their corresponding declarative controls; and (b) Experiment 1b, which consisted of a negation test, to measure the backgroundedness of sentence complements of those verbs where extraction appeared.

This experiment tested all three previously existing accounts. The BCI account predicts a negative correlation between the backgroundedness of the extraction domain and the acceptability of the wh-questions (A&G, 2008). The syntactic accounts (e.g., Snyder, 1992) predict that all the wh-questions formed by factive and manner-of-speaking verbs should be less acceptable than all the bridge verb extractions. The semantic accounts (e.g., Kiparsky & Kiparsky, 1971) predict that all the factive wh-questions should be less acceptable than all the bridge and manner-of-speaking verb extractions.

2.2.1 Participants

180 subjects participated in this experiment via Amazon Mechanical Turk. 120 participants rated the acceptability of wh-questions and declarative clauses (Experiment 1a); another 60 subjects completed the negation task (Experiment 1b). The experiment was only visible to people who had a U.S. IP address. Participants were asked to indicate their native language, but payment did not depend on their answer to this question.

2.2.2 Design and Materials

The acceptability and negation tasks were constructed using 24 sentence complement verbs of the 3 categories, as listed in (11).

(11) a. Bridge verbs: **say, decide, think, believe**, feel, hope, claim, report, declare

b. Factive verbs: **know, realize, remember, notice,** discover, forget

c. Manner-of-speaking verbs: **whisper, stammer, mumble, mutter,** shout, yell, scream, murmur, whine

Verbs in bold were those tested in A&G (2008). The labeling of a verb as 'bridge' was obtained from previous literature, such as Erteschik-Shir (1973, 1979, 2007), Snyder (1992), Ambridge and Goldberg (2008), and Goldberg (2013, 2016). In the acceptability task, wh-questions and their corresponding declarative sentences were designed as in (12a) and (12b) respectively. 96 pairs of wh-questions and declaratives were constructed, and each of the 24 tested verbs in (11) formed 4 pairs. In each pair of wh-question and declarative control, NP1 and NP2 were common names, V1 came from (11), and V2 was the past tense form of one of 25 frequently used verbs (*like, eat, buy, build, cook, destroy, dislike, drink, draw, fix, find, know, learn, lose, make, mention, need, see, sell, steal, take, teach, throw, want, write*). To reduce the possibility of semantic plausibility confounds, we used 'something' instead of a specific NP as the embedded object, as shown in (12b).

(12) a. What did [NP1] [V1] [[that] [NP2] [V2]]?

e.g., What did Susan know that Anthony liked?

b. [NP1] [V1] [[that] [NP2] [V2+something]]

e.g., Susan knew that Anthony liked something

The 96 pairs were split across 2 lists: each list contained 2 declaratives and 2 wh-questions per verb. Each participant saw 96 sentences (from one list) in a random order. They were asked to rate how natural each sentence was using a rating scale from 1 (extremely unnatural) to 5 (extremely natural). Each sentence was followed by a comprehension question about the content of the sentence to check if participants were paying attention to the task (e.g., 'Does this sentence mention Andy?').

In the negation-test task, each trial included a negated complex sentence (13a) and a negated simple sentence (13b) which was the negated version of the sentence complement in (13a).

(13) a. [NP1] didn't [V1] [that] [NP2] [V2+Appropriate NP]
e.g., Susan didn't know that Anthony liked the cake.
b. [NP2] didn't [V2+Appropriate NP]
e.g., Anthony didn't like the cake.

Participants were asked to rate how true they thought the second sentence was, given the first sentence, with a scale from 1 (false) to 5 (true). A&G (2008) proposed that these negation scores should reflect how "backgrounded" the information in the sentence complement is.

2.2.3 Results

In all the experiments reported here, data from participants who did not self-report as native speakers of American English or didn't answer all the comprehension questions with at least 85% accuracy were excluded. Responses from 116 participants in the acceptability task and 49 participants in the negation task were analyzed.

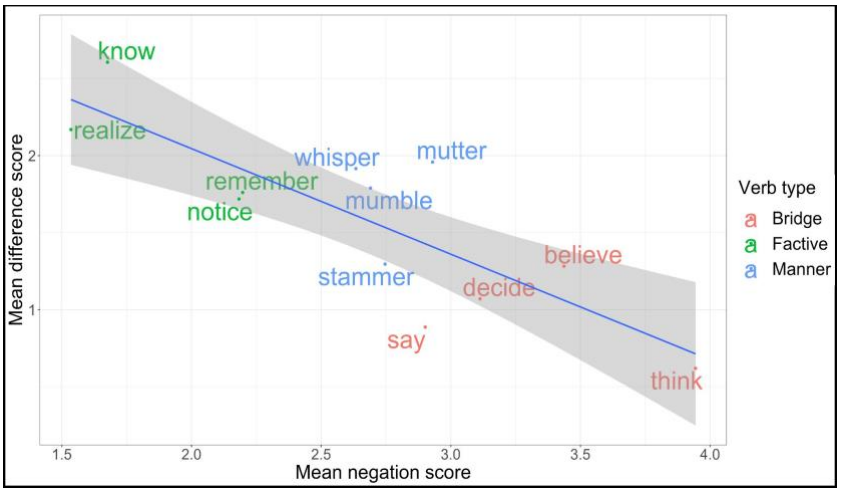
A. Results of the negation-acceptability analysis of A&G (2008)

In A&G (2008), 71 participants were recruited for both tasks. The authors calculated the difference scores between the ratings of wh-questions and declarative clauses as the measurement for acceptability of those wh-questions, and they found a strong Pearson correlation between these difference scores and the negation scores, calculated on each verb ($r=-0.83$, $p<0.001$; see Figure 3a). We applied an analogous analysis to our data. The obtained correlation in our data was in the same direction as in A&G (2008), but the effect was smaller and non-significant both in the 12 verbs they tested ($r=-0.40$, $p=0.20$; see Figure 3b) and in the full set of 24 verbs ($r=-0.31$, $p=0.13$; see Figure 3c)[7]. Experiments in the original study were conducted on a 7-point Likert scale, while ours are on a 5-point scale. Since people were mostly using the top of the scale (3-5 in ours, probably 4-7 in the original study), the difference scores are smaller in our study.

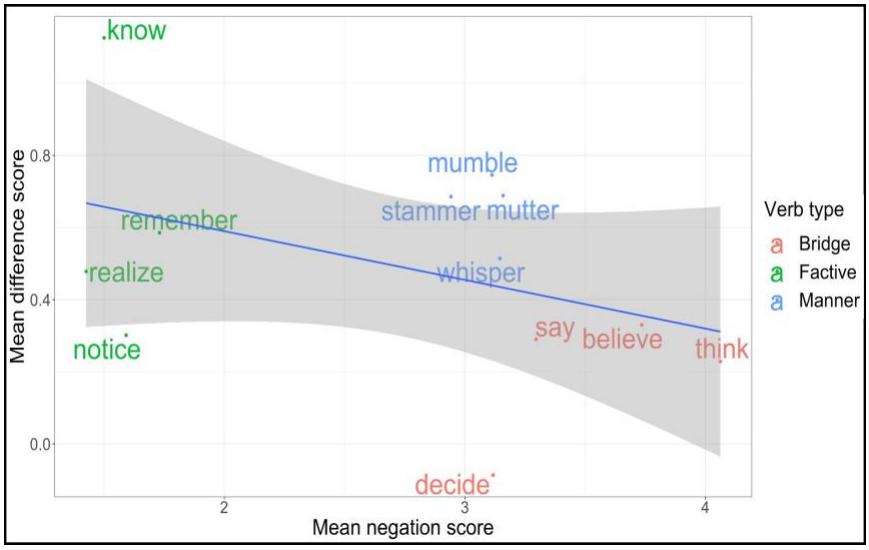
The lower correlations that we observed appear to be derived from at least two sources: first, manner-of-speaking verbs have highly variable difference scores, but very similar negation scores; and second, factive and manner-of-speaking verbs have overall similar difference scores but very different negation scores. Given the larger sample size (i.e. more tested verbs), it is likely that our dataset provides a more accurate estimate of the effect size.

[7]Note: This was not a direct replication. For example, in contrast to A&G (2008), acceptability and negation scores were collected on different subjects.

3a. Results from A&G (2008) (12 verbs on a 7-point Likert scale).



3b. Results from 12 tested verbs in the present study (5-point Likert scale).



3c. Results from all the tested verbs in the present study (24 verbs on a 5-point Likert scale).

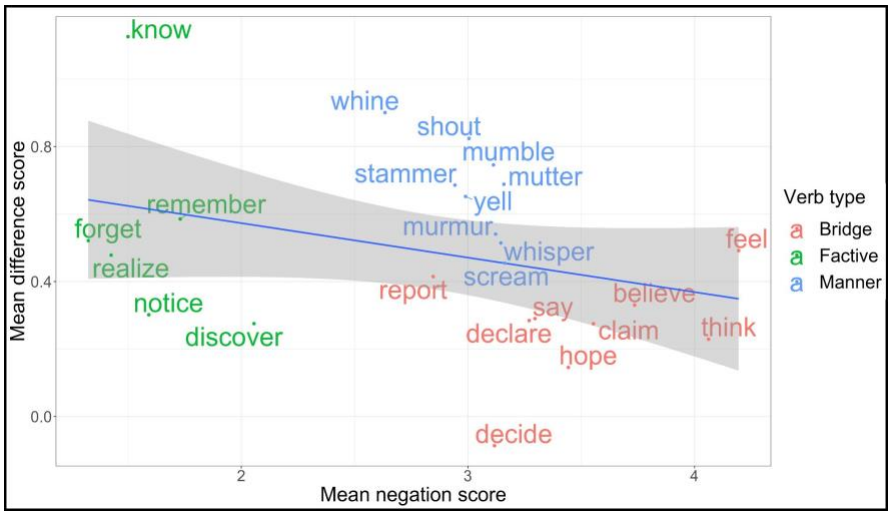


Figure 3. Correlation between mean difference scores and mean negation test scores by verb in A&G (2008) and in the present study (Experiment 1).

In addition, we found large overlap between acceptabilities for factive and bridge wh-questions (Figure 4), contradicting the syntactic and semantic accounts, which predict non-overlapping acceptability between factive and bridge wh-questions given their distinct covert deep structures (Figures 2a and 2b). Note that the acceptability of manner-of-speaking verb wh-questions was more similar to the factive verb wh-questions than the bridge verb wh-questions, which further challenges the semantic accounts, because they group bridge and manner-of-speaking verbs together, since only factive verbs take presuppositions (Figure 2b). Our results are consistent with those of Kothari (2008) who showed that there is no categorical acceptability distinction between extraction across manner-of-speaking and non-manner-of-speaking verbs.

Following reviewers' suggestions, we conducted two further analyses of the BCI account, which we present in Appendix A: (i) ordinal regression analyses were applied to our collected data to further test the discourse BCI account; (ii) a Bayes factor analysis to weigh the evidence for and against the

presence of the discourse BCI effect (i.e., an interaction between sentence type and negation scores, in this case). Results of these analyses suggested that there was no robust evidence for the discourse BCI effect in our dataset.

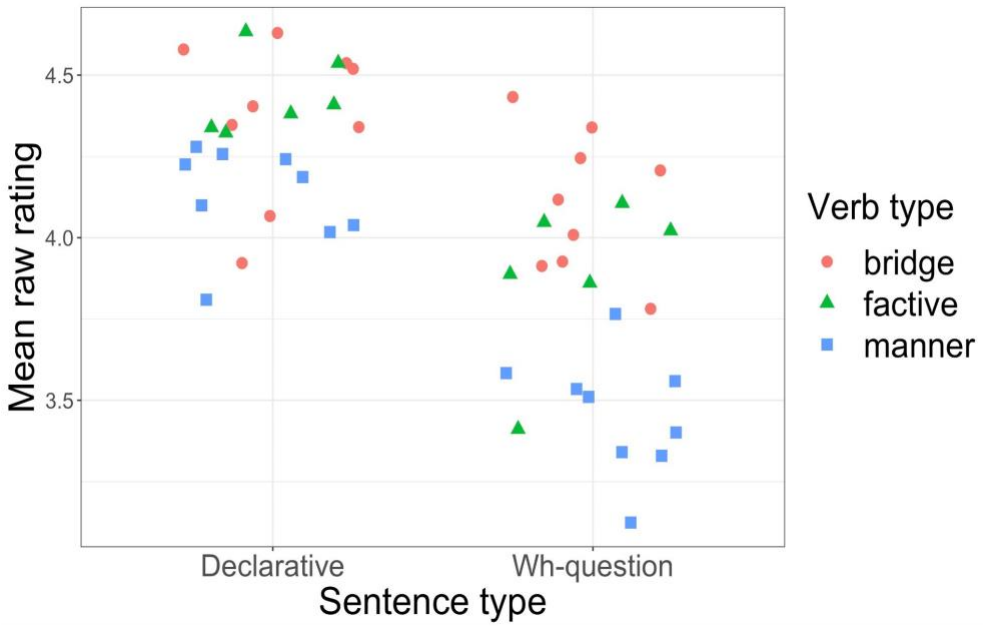


Figure 4. Mean ratings of wh-questions and declarative clauses by verb in Experiment 1, jittered for visualization purposes, for comparison with predictions of the syntactic and semantic accounts in Figures 2a and 2b.

In sum, we didn't find strong supportive evidence for the BCI account. Furthermore, our findings were not in line with the previous syntactic or semantic approaches to explaining these islands.

B. The verb-frame frequency account and results of post hoc analyses

We also evaluated our simpler hypothesis: the *verb-frame frequency hypothesis*, restated below. We collected the frequencies of the 24 verbs followed by the complementizer 'that' from the Google books corpus (since the year 2000) as a proxy for relative verb-frame frequency. The 24 words were labeled as verbs and searched with all the possible tense and aspects in Google books.[8]

[8] We also counted the frequencies of those verbs taking sentence complements in two parsed English corpora: the Wall Street

The Verb-frame Frequency Hypothesis:

The acceptability of a sentence is best captured by two independent, separate effects: (i) the frequency or the type of the construction (e.g., wh-questions vs. declaratives) and (ii) the frequency of the verb head-structure, P(matrix verb, sentence complement) = P (matrix verb) * P (sentence complement | matrix verb).

Because the outcomes were Likert scale ratings, we applied mixed-effects ordinal regression in the *ordinal* package in R. Though it is common in studies of the island phenomena to apply linear models to Likert scale rating data, this method might lead to spurious results if the data are skewed toward one end of the scale (e.g., Liddell & Kruschke, 2018). In the present dataset, most (74.6%) of the responses are 4 or 5, as in Figure 5[9]. Moreover, treating Likert scale rating data as a metric scale assumes there are equal distances between the ordinal ratings (1-5), which is not necessarily the case. For instance, the true acceptability difference between 3 and 4 may not be the same as that between 4 and 5, though the metric difference is 1 in both cases.

Journal and Brown corpus (both in the Penn Treebank). There were fewer than 5 instances of the low-frequency verbs co-occurring with clausal complements, which consisted of many of the manner-of-speaking verbs (e.g, 'whisper'). Consequently, we used frequencies estimated via the Google books corpus. In addition, for the higher frequency verbs, the log-transformed frequencies of those verbs taking clausal complements in the Wall Street Journal and Brown corpus are highly correlated with Google books frequencies ($r = 0.9$, $p<0.001$). See the results section of Experiment 2 for more details.

[9] Over 50% responses of the declaratives and around half (43.9%) of all the responses are distributed at the ceiling of the whole scale, rating 5. The responses of rating 4 and 5 occupy 74.6% of all the responses, while only 1.8% of the responses are the lowest rating 1.

Figure 5. The distribution of acceptability ratings on the 5-point Likert scale by sentence type in Experiment 1.

We entered *sentence type* (declarative vs. wh-question), *log-transformed verb-frame frequency*, and their *interaction* as the predictors. The model was fitted with the maximum random effect structure which contained random intercepts for *subjects* and *verbs* as well as by-subject slopes for the effects of *sentence type*, *frequency*, and their *interaction* and by-verb *sentence type* slopes. Consistent with the verb-frame frequency hypothesis, log-transformed verb-frame frequency had a significant impact on the acceptability ratings (β=0.50, Z =5.89, $p<0.001$). Wh-questions were significantly less acceptable than declaratives (β=-1.40, Z =-7.04, $p<0.001$). The interaction of sentence type and verb-frame frequency was not a significant predictor ($p>0.08$) of acceptability ratings[10].

[10] We applied an ordinal regression analysis to the data from A&G (2008) (which were kindly supplied by Ben Ambridge), to see whether the previously observed significant interaction between sentence type and negation score was due to the use of a linear model on ordinal data. The results -- provided in Appendix A -- showed that both linear and ordinal regressions applied to the dataset in A&G (2008) yielded a significant interaction effect. Hence there seem to be differences between the results from our data set and those from A&G (2008), perhaps due to the greater variety of materials in our set, or some other difference between the experimental materials and/or fillers.

Due to concerns about the interpretation of skewed ordinal data, in an exploratory analysis, we converted the 5-point scale responses into binary outcomes (acceptable = 1, unacceptable = 0). Two transformations were used and analyzed: (i) transformation of rating 1-2 to 0 and rating 3-5 to 1; or (ii) transformation of rating 1-3 into 0 and rating 4-5 into 1. Mixed-effects logistic regressions in the *lme4* package in *R* with the same fixed and random effects as the ordinal regression were applied to the binarized rating responses (one for each way of binarizing the data). Results from the two models were qualitatively similar. For instance, the model fit on data with transformation (i) showed that both sentence type (β=-2.10, Z =-6.68, $p<0.001$) and frequency (β=0.45, Z =3.85, $p<0.001$) were significant predictors of acceptability. The interaction of frequency and sentence type had no significant impact on the outcome (β=-0.09, Z=-0.44, p=0.66) as shown in Figure 6.[11] The full table of results of all the regression analyses reported in the main text of this paper are attached in Appendix C.

[11]A possible outlier for the frequency account is the verb 'know' (bottom right on in Figure 6), which is low in acceptability despite its high frequency. We discuss this issue following Experiment 3.

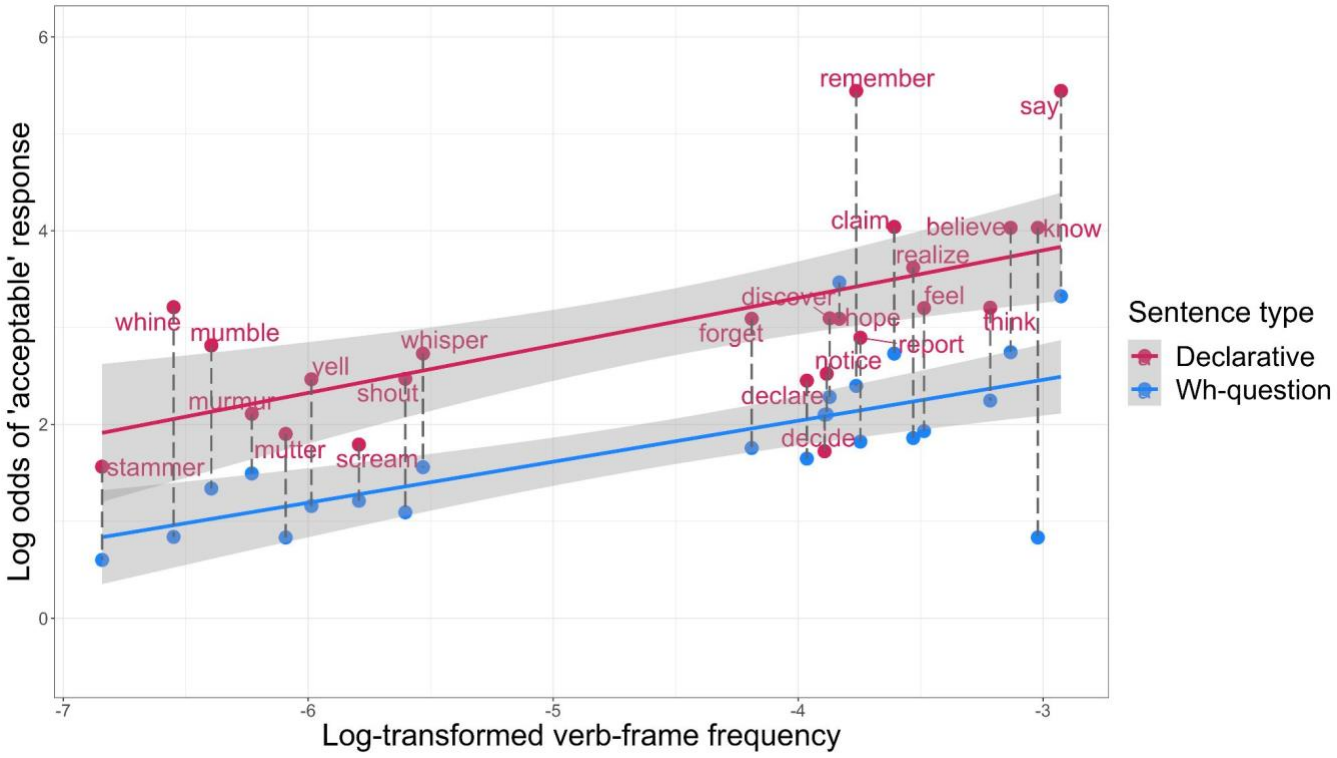


Figure 6. Results of Experiment 1: converted log odds of 'acceptable' response for wh-questions and declarative clauses (transformation of rating 1-2 to 0 and rating 3-5 to 1) against log-transformed verb-frame frequencies by verb. The dashed lines link the two instances of each verb.

We also performed model comparison between models fit based on the discourse and the frequency accounts. The results showed that the model of verb-frame frequency account is favored in terms of Bayesian Information Criterion (BIC). See Section III in Appendix A for more details.

2.2.4 Discussion

Contrary to the three previous accounts of factive and manner-of-speaking islands (Ambridge & Goldberg, 2008; Kiparsky & Kiparsky, 1971; Snyder, 1992), we found no robust evidence for factors that solely influence wh-questions but not declaratives. The previous quantitative evaluation of these islands had only 12 verbs (Ambridge & Goldberg, 2008). It is possible that the larger sample size of verbs in our dataset provides a more accurate estimate of the effect size.

Our exploratory analyses provide initial support for the verb-frame frequency hypothesis. Sentence type and verb-frame frequency have additive and independent effects on the acceptability of wh-questions and declaratives. In Experiment 2, we sought to replicate and extend these findings using a larger set of verbs and a binary dependent measure.

2.3 Experiment 2: Wh-questions with 48 Verbs

The goal of Experiment 2 was to test the frequency account with more matrix verbs beyond the three categories (bridge, factive, manner-of-speaking). The verb-frame frequency hypothesis predicts that the verbs that frequently take sentence complements should be more acceptable in wh-questions and declaratives, regardless of the verb category. The syntactic and semantic accounts discussed in Experiment 1 cannot explain extraction across verbs beyond the three categories. Previous theories all predict a significant interaction between verb-frame frequency and construction type (declarative vs. wh-question), whereas the frequency account predicts no such interaction.

Given that most participants in Experiment 1 were not using most of the 5-point Likert scales, we used a forced-choice binary acceptability judgment task in this experiment. Results from previous studies (e.g., Weskott & Fanselow, 2011; Sprouse et al., 2013) have shown that different measurements (e.g., Likert scales, binary scale, or magnitude estimation) lead to very similar results, with the consequence that changing this detail of the method should have little effect on the results.[12]

[12] Indeed, Experiment 3 was run in two variants -- forced-choice binary acceptability judgment, and a 5-point acceptability scale (Experiment 4) -- and the results were remarkably similar across the two.(For details, see Experiments 3 and 4).

2.3.1 Participants

120 people participated via MTurk. The experiment was only visible to people who had a U.S. IP address.

2.3.2 Design and Materials

The design was similar to Experiment 1a, with 48 verbs that could take sentence complements. The materials included 8 verbs from each of the three categories (bridge, factive,and manner-of-speaking) and another 24 verbs outside the three categories, as listed in (14). The 24 'other' verbs were not clearly categorized in the previous literature. Given that the notion of 'bridge' is undefined, the concept of 'factivity' is gradient, and there is no exhaustive list of manner-of-speaking verbs, we cannot rule out the possibility that some of these 24 verbs may fall within the three categories, according to certain researchers' guidelines. Critically, the major predictor for acceptability of wh-questions/declaratives is verb-frame frequency, not which category each verb belongs to.

(14) **Matrix verbs**:

Bridge (8): feel, say, believe, hope, think, report, declare, claim,

Factive (8): know, remember, realize, notice, discover, forget, learn, hate

Manner (8): whisper, mumble, murmur, mutter, whine, shout, yell, scream

Other (24): hear, recall, blab, conjecture, conceal, proclaim, hint, remark, infer, confirm, deny, guess, confide, maintain, testify, reveal, suspect, verify, prove, insist, guarantee, presume, hypothesize, complain

Wh-questions and declaratives were constructed for the 48 matrix verbs with 6 items for each verb (288 items in total). A sample item is given in (15). To keep items as plausible as possible, we used two kinds

of verbs in the most embedded position: action (e.g., *bought, wrote*) and mental (e.g., *wanted, liked*). 42 out of the 48 matrix verbs were paired with the 6 action embedded verbs in (16a). The two mental matrix verbs (*feel, insist*) were matched with 6 mental embedded verbs (16b), because these only make sense with mental embedded verbs. The 4 remaining matrix verbs (*hope, guarantee, presume, hypothesize)* worked well with both kinds of embedded verbs, so we selected some from each set for each of these verbs. A set of examples of tested wh-questions is provided in (17).

(15) a. What did [NP1] [VERB1] [[that] [NP2] [VERB2]]?

(e.g., What$_i$ did Susan know that Anthony bought __$_i$?)

b. [NP1] [VERB1] [[that] [NP2] [VERB2+something]]

(e.g., Susan knew that Anthony bought something)

(16) Embedded verbs[13]

a. Action (6): bought, wrote, sold, took, stole, broke

b. Mental (6): wanted, liked, disliked, preferred, needed, loved

(17) a. What did Melissa say that Eric wrote?

b. What did Amanda feel that Jason liked?

c. What did Linda insist that John wanted?

[13] We wanted to use a small set of embedded verbs for the 48 tested matrix verbs, so that random meaning differences in the embedded clauses would be reduced. While most of the tested matrix verbs can be paired with transitive verbs denoting action such as 'buy' to form a plausible sentence (e.g., 'What did John say/confirm that Mary bought?'), some verbs such as 'feel' cannot always be paired with action verbs as in (16a) (e.g., ???'What did John feel that Mary bought?'). For such verbs, we used the set of "mental" verbs in (16b) (e.g., 'What did John feel that Mary liked?').

As in Experiment 1a, participants were assigned to 1 of 2 lists made up of 3 declaratives and 3 wh-questions for each of the 48 verbs. Each participant saw 288 sentences in a random order. Participants were asked to rate each sentence using a binary scale (acceptable vs. unacceptable) based on how natural they thought the sentence was. Each sentence was also followed by a comprehension question.

2.3.3 Results

Data from subjects who were not native speakers of American English or who did not answer all the comprehension questions with at least 85% accuracy were excluded. Responses from 110 participants were analyzed.

To check the validity of the verb-frame frequencies that we had estimated via the Google books corpus, we obtained frequencies of all verbs in our 48 tested verbs in the parsed Wall Street Journal and Brown corpus from the Penn Treebank which were followed by a sentential complement (with or without the complementizer 'that') and had at least 5 instances in the corpora. This resulted in 19 verbs. These Log-transformed verb-frame frequencies (P (verb, sentence complement)) were highly correlated with the Google books measures ($r = 0.9$, $t = 8.48$, $p < 0.0001$), suggesting that the verb-frame frequencies obtained from Google books are valid.

Acceptability judgments were analyzed with a mixed-effects logistic regression using the *lme4* package in *R. Sentence type (declarative vs. wh-question)*, *log-transformed frequency of the verb frame* and their *interaction* were entered as predictors. The model was fit with the maximum random effect structure which contained random by-*subject* and by-*verb* intercepts as well as slopes for *sentence type*frequency* by-subject and slopes for *sentence type* by-verb. The log-odds of an 'acceptable' response for declaratives and wh-questions for a given verb-frame frequency are plotted in Figure 7.

The results supported the verb-frame frequency hypothesis. Wh-questions and declaratives formed by verbs of higher verb-frame frequency were significantly more acceptable (β=0.59, z=3.95, p<0.001). There was also a significant main effect of sentence type: declaratives were rated more acceptable than wh-questions (β=-2.45, z=-7.88, p<0.001). No interaction was found (p>0.4). If anything, Figure 7 shows a pattern resembling a numeric interaction in the opposite direction. That is, a theory that predicted an interaction would predict the effect of frequency would have a steeper slope for wh-questions than declaratives.

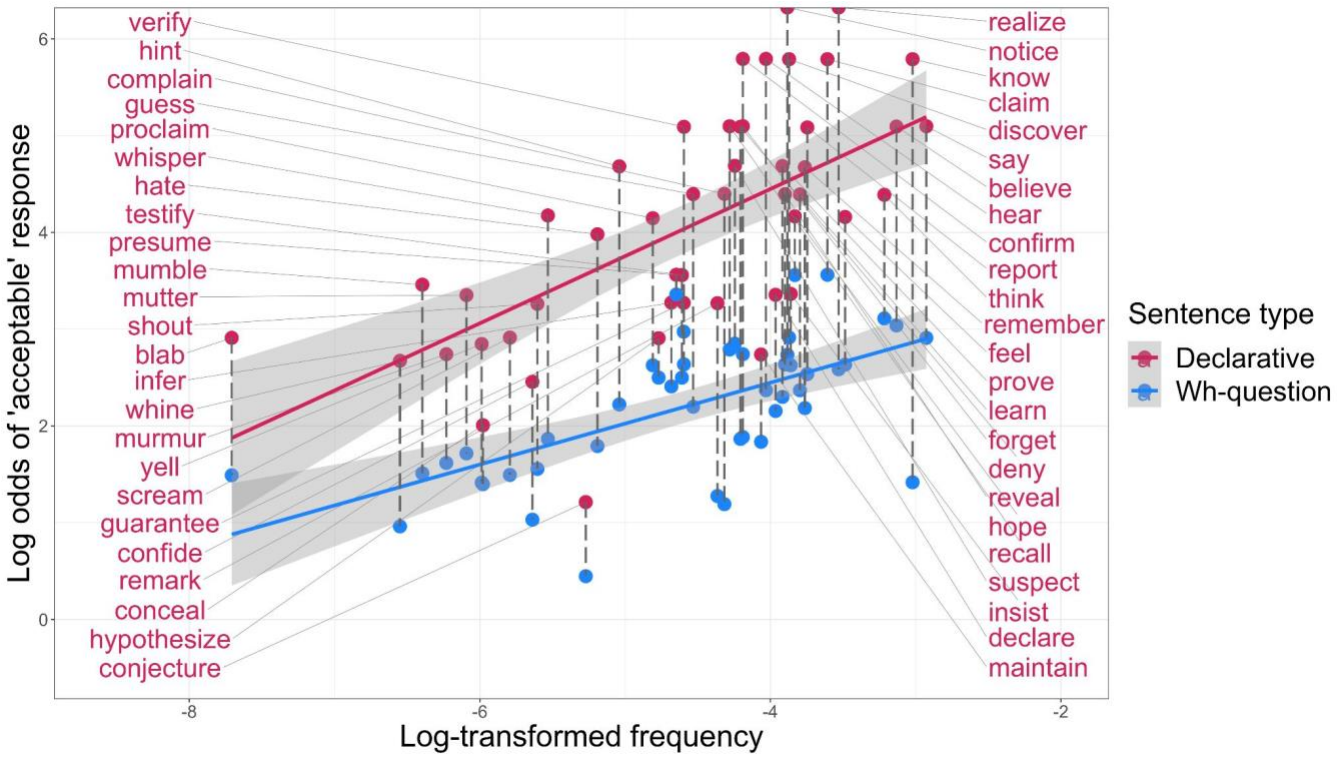


Figure 7. Results of Experiment 2: log-odds of 'acceptable' response for wh-questions and declarative clauses against log-transformed frequencies by verb (48 verbs). The dashed lines link the two instances of each verb.

As Experiment 1 evaluated a subset of the verbs examined in Experiment 2, we investigated the stability of ratings across these two experiments. To do so, we extracted the 22 verbs that were investigated in

both Experiments 1 and 2, and calculated the mean rating (Experiment 1) and the proportion of 'acceptable' responses (Experiment 2) for declaratives and wh-questions for each of these 22 verbs. This analysis revealed that mean ratings from Experiment 1 were highly correlated with the proportion of 'acceptable' responses in Experiment 2 ($r = 0.92$, $t = 15.9$, $p<0.0001$).

2.3.4 Discussion

In Experiment 2, we replicated and extended Experiment 1, and showed that the verb-frame frequency account provides a better explanation for wh-question and declarative acceptability than previous accounts because it can explain within-verb category variance and overlap across verb categories. Further, it can capture acceptability of wh-questions and declaratives formed by verbs outside the 3 categories.

In a followup to our work, Richter & Chaves (2020), performed an acceptability study on wh-questions formed by 75 sentence complement verbs from 5 categories: (a) 15 factives, (b) 15 manner of speaking, and 45 other verbs with 3 frequency levels - (c) 15 low frequency, (d) 15 medium frequency, and (e) 15 high frequency. They found a robust effect of verb subcategorization frequency on wh-question acceptability, consistent with our findings.[14]

[14] Richter & Chaves (2020) suggest that a weakness of our work is that once the interaction between verb subcategorization frequency and verb type is entered into the model (in addition to these two main effects), the effect of verb subcategorization disappears. The authors argued that these results suggest that verbs of different types are distributed very differently with respect to subcategorization frequency, which they suggest challenges the breadth of a verb-frame frequency-based account. There are several issues with Richter & Chaves' methods that make their critique hard to interpret. First, the categorization of verb type -- which is crucial to the interpretation of this model -- has no empirical basis. As we have discussed, there is no independent empirical test that can divide these verbs into the categories bridge, factive, manner, and other. The low/middle/high frequency distinction is also arbitrary. Second, Richter & Chaves used raw frequencies in their model, rather than log-transformed frequencies, in contrast to what languages researchers standardly use. Raw frequencies are highly skewed, and cannot pick up variance for lower frequency verbs. Third, the corpora that they used were VALEX and COCA, which are much smaller than the Google books corpus that we used. The advantage of the larger corpus is that its estimates for lower frequency verbs are much better (which becomes more useful if one analyzes log frequency). Finally, while we had two conditions for each verb -- wh-question and declarative -- each with ratings in our experimental design and statistical model, Richter & Chaves only included one condition for each verb: the wh-question version. They performed a separate "control" experiment for the declarative versions, and entered the mean of those values for each verb as a random intercept in

In Experiment 3, we sought to evaluate the verb-frame frequency account in another filler-gap construction, the cleft structure.

2.4 Experiment 3: Cleft Structures

Experiment 3 aimed to further test the verb-frame frequency account on another filler-gap construction, the cleft structure. We chose to test cleft structures rather than relative clauses, because clefts have fewer content words compared to relative clauses and therefore introduce less additional noise when compared with declaratives. The verb-frame frequency account predicts that frequency plays the same role in the acceptability of both declaratives and clefts. Cleft structures should be rated less acceptable than declaratives, perhaps because people produce more declaratives than clefts (or are perhaps due to other other cognitive constraints, such as working memory demands (e.g., Gibson, 1998)).

2.4.1 Participants

Data from 120 participants were collected via MTurk. The experiment was only visible to people who have a U.S. IP address.

2.4.2 Design and Materials

their model. This is an odd way of modeling the data: Given that we show that declarative and wh-question ratings are similarly influenced by verb frequency and verb type (however categorized), having declarative rating as a random intercept will likely cause collinearity issues. The variance in the dependent variable -- wh-question ratings -- that is supposed to be captured by the fixed effects -- verb type and frequency -- might then be wrongly attributed to the random intercept, from the declarative ratings.

Cleft structures and their corresponding declarative sentences were designed as in (18a) and (18b) respectively. 96 pairs of clefts and declaratives were constructed. We tested the same 24 verbs as in Experiment 1 in (11). Each of the 24 tested verbs formed 4 pairs as in Experiment 1a.

(18) a. It was [NP3][that][[NP1] [VERB1] [that][[NP2] [VERB2]]

(e.g., It was the pie that Angela mumbled that Kevin liked)

b. [NP1] [VERB1] [that] [[NP2] [VERB2+NP3]]

(e.g., Angela mumbled that Kevin liked the pie.)

The 96 pairs were split across 2 lists: each list contained 2 declaratives and 2 cleft structures per verb. Each participant saw 96 sentences (from 1 list) in a random order. Participants were asked to rate each sentence with a binary rating scale. Each sentence was followed by a comprehension question (e.g., 'Does this sentence mention an apple?')[15].

2.4.3 Results

We excluded data from subjects who did not identify as native speakers of American English or who did not answer all the comprehension questions with at least 85% accuracy. Responses from 104 participants were analyzed.

Acceptability responses were analyzed as in Experiment 2. Sentences with higher frequency verb frames were significantly more acceptable (β=1.24, z=2.4, $p < 0.02$) and cleft structures were less likely to be

[15] We didn't include fillers in the experiments. There were many items, and each list contained at least 96 sentences, so adding fillers would make the list too long for each participant. In addition, other experiments have shown very similar results with and without fillers for acceptability rating tasks (Gibson et al., 2012).

acceptable (β=-10.7, z= -4.94, p < 0.001). The interaction of sentence type and frequency was not significant (β=-0.87, z=-0.84, p=0.4) (Figure 8). These data are best explained by positing that verb frame frequency and extraction have independent, additive effects in log-odds space, as predicted by the verb-frame frequency account.

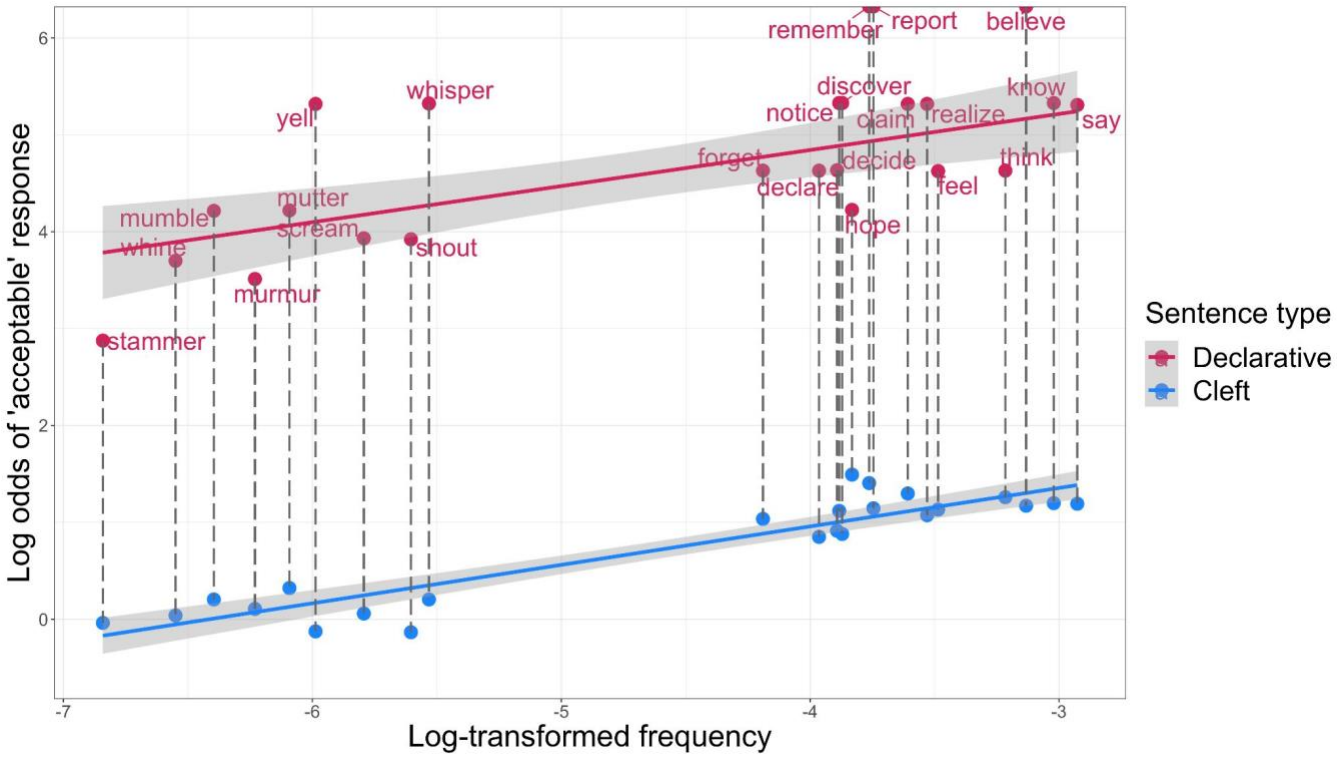


Figure 8. Results of Experiment 3: Log-odds of 'acceptable' response for clefts and declaratives against log-transformed frequencies (24 verbs). The dashed lines link the two instances of each verb.

We also ran a 5-point Likert scale version of this experiment and the results were qualitatively the same. When analyzed using an ordinal model, we found main effects of sentence type (declarative vs. cleft) and verb-frame frequency, but no interaction. See Experiment 4 in Appendix B for details.[16]

[16] The results of Experiment 4 also showed that the application of ordinal and linear regressions to the same dataset can lead to different results. When these data were analyzed using a linear model, a significant interaction between sentence type and verb-frame frequency was observed, as in Appendix B, which suggests applying linear models to ordinal data can lead to false positives - a spurious interaction (Liddell & Kruschke, 2018).

As discussed in the introduction to Experiment 2, it is unsurprising that these two slightly different methods -- binary judgement vs. 5-point acceptability scale -- result in similar statistical conclusions, because different measurements (e.g., Likert scales, binary scale, or magnitude estimation) tend to lead to similar results (e.g., Weskott & Fanselow, 2011; Sprouse et al., 2013). We consider the 5-point Likert scale version of this experiment a replication.[17]

2.4.4 Discussion

The results of Experiment 3 provide further evidence for the verb-frame frequency account with another type of filler-gap construction – cleft structures. Like in Experiments 1 and 2, we found that materials using the filler-gap construction -- the cleft -- were rated as less acceptable than their declarative counterparts and materials with higher verb frame frequencies were rated as more acceptable.

A visual comparison of results from Experiments 2 and 3 suggests that clefts may have received lower ratings than wh-questions, but a statistical comparison is difficult to make between these experiments. If this difference between clefts and wh-questions is real, it could come from several sources: clefts as a construction are rarer than wh-questions; alternatively, it could be that a null context (as in this experiment) simply doesn't license a cleft as well as a wh-question. Consequently, we are cautious not to over-interpret these rating differences.

[17] Although there is a tendency to think that a multi-point scale will give more precise item measures than a binary judgement task, it turns out that this is not the case. This is plausibly because people can't remember what rating they gave to more than a few items, so internal consistency is difficult across items, except when simply judging materials independently of each other. Consequently, the best way to get good item estimates is through many samples, across participants, not through a more precise measure for each participant.

Testing cleft structures also allowed us to evaluate whether a potential outlier to the frequency account in Experiments 1 and 2 - the verb 'know' - might be explained by pragmatic factors, having to do with the meaning of the wh-question construction. The verb 'know' is a very frequent verb, and yet it is not very acceptable in the wh-question forms in Experiments 1 and 2 (it is the bottom right dot in each of Figures 6 and 7). We speculate that the idiosyncratic behavior of 'know' in wh-questions may be due to pragmatic factors in wh-questions: a question is a request for knowledge but the verb 'know' has its primary conventionalized meaning that the subject has the knowledge indicated in the embedded sentence. Thus, it may be somewhat incoherent for the meaning of the wh-question to contradict the primary meaning of the verb "know". This pragmatic hypothesis does not apply to other (factive) verbs. 'Know', unlike other (factive) verbs, does not have additional meaning other than having the knowledge of the event. But other (factive) verbs have additional conventionalized meaning, so that the meaning of the wh-question does not contradict the primary meaning of the embedding verb. For example, the meaning of "forget" focuses on 'failing to remember' rather than 'having the knowledge', so there is no direct contradiction with the meaning of a wh-question. The pragmatic hypothesis predicts that 'know' should be acceptable in other filler-gap constructions whose meaning is not requesting knowledge. In line with this speculation, we found that 'know' is not an outlier for the frequency account in the cleft structure (Figure 8). Further work is needed to evaluate how "know" is used across constructions to see if this kind of cross-construction usage idea applies more generally (c.f. Abeillé et al., 2020).

2.5 General Discussion

The results of all three experiments show that verb-frame frequency is a determining factor for the acceptability of filler-gap constructions formed by various sentence complement verbs, including factive and manner-of-speaking verbs. Experiment 1 consisted of a replication and extension of Ambridge and

Goldberg (2008), with 24 sentence complement verbs across bridge, factive, and manner-of-speaking verbs. We found that the existing discourse, syntactic and semantic accounts could not explain the pattern of data that we observed. We therefore proposed and tested the verb-frame frequency account. The results of Experiment 1 were as predicted by such an account: there were main effects of verb-frame frequency and construction type/frequency, with no interaction. Experiment 2 was designed to further test the verb-frame frequency account with a broader set of 48 sentence complement verbs beyond the three initial categories. The results confirmed the verb-frame frequency account - verbs of higher verb-frame frequency were significantly more acceptable, and declaratives were more acceptable than wh-questions, with no interaction between the two. In Experiment 3, we further tested the frequency account on cleft structures, another type of filler-gap construction. The results provided further support for the frequency account: Two main effects, verb-frame frequency and construction type, were found, with no interaction between the two. Taken together, these results indicate that verb-frame frequency robustly predicts acceptability ratings in sentences with long-distance dependencies. This account is favored by Occam's Razor, as it has few parameters: verb-frame frequency and sentence type. We leave it to future research to explain variance that remains unaccounted for by this account.

2.5.1 Relation to theories of sentence processing

One may ask whether frequency is the cause of unnaturalness in filler-gap constructions, or whether usage frequencies are merely a reflection of discourse/meaning/structure factors which are the true causes of unacceptability. First, frequencies in natural language might come from many sources, including but not limited to the factors we have evaluated. For example, perhaps some verbs take sentence complements more frequently because of the typicality of the way of speaking: *saying* something (in a normal tone of voice) is more common than *whispering* or *shouting* or other manners of speaking. This would partially explain the high frequency of '*say* sentence-complement' compared to

'*whisper* sentence-complement', for example. Second, while frequencies may be underlyingly caused by such hidden factors, the tight fit between acceptability ratings and frequencies suggests that frequency may form a causal bottleneck mediating the effect of these factors on acceptability ratings. That is, we propose that discourse/semantic/structural factors might give rise to frequency distributions, and frequency distributions give rise to acceptability ratings. Thus, discourse/meaning/syntax and acceptability judgments are conditionally independent given knowledge of frequency. This logic is similar to the idea of the 'surprisal bottleneck' in psycholinguistics (Levy, 2008; Smith & Levy, 2013), which holds that syntactic and semantic factors cause processing difficulty only by modulating the probabilities of words in context.

An open question for this research program is why it is that the matrix verb-frame frequency seems to have a particularly strong effect on acceptability in these phenomena, but not the frequency of all of the words / constituents equally. A partial answer to this question is that the verb is typically considered as the head of an event structure, on which other constituents depend. It is therefore perhaps unsurprising that lexico-syntactic information carried by verbs can have an important effect on sentence acceptability. That being said, manipulating other parts of the sentence may also lead to differences in acceptability. For instance, 'What did the *teacher* say that the boy wrote?' may sound more natural than 'What did the *schoolmistress* say that the boy wrote?'. But frequency changes in these constituents seem to result in relatively minor differences. Of related interest is the observation that the matrix verb seems to play a larger role in acceptability than the embedded verb. For example, 'What did John *say* that Mary *muttered*?' sounds more acceptable than 'What did John *mutter* that Mary *said*?', though these two sentences contain identical verbs. We leave these puzzles to future research to resolve.

2.5.2 Learnability of islands

The finding that the acceptability of wh-questions is highly correlated with verb-frame frequency suggests that the unacceptability of certain filler-gap constructions is modulated by exposure, and is therefore learnable, which challenges the traditional (Universal Grammar) view that the unacceptability of filler-gap constructions is not learnable and must to be innate (Chomsky, 1986). Although direct negative evidence is missing especially for such complex structures, children may draw statistical inferences from the input and regard the absence of a certain input (e.g., a type of extraction) as evidence of its oddness (rendering it unacceptable) (cf. Hsu & Griffiths, 2016; Kidd, Lieven & Tomasello, 2010; Navarro, Dry & Lee, 2012; Voorspoels, Perfors, Ransom & Storms, 2015; Xu & Tenenbaum, 2007).

2.5.3 Connection to syntactic theories

Though we did not find support for syntactic accounts for extraction difficulty in factive and manner-of-speaking structures, this project does not deny the importance of syntactic structure in language processing and learning. Indeed, by considering alternatives to covert structures that are not supported by independent empirical evidence and proposing the same structure for all the sentence complement verbs, we may in fact reach a more efficient and simpler syntactic framework (c.f., Culicover & Jackendoff, 2005).

2.6 Materials:

The data and materials are publicly available at https://osf.io/2ydqc/.

Chapter 3: [Paper 2] Logophoric Chinese reflexives *ziji* and *taziji*

3.1 Introduction

Reflexives across languages, though sharing some similarities, demonstrate diverse referential properties, which challenge Condition A from Chomsky's (1981) Binding theory (Binding Condition A = Condition A). In particular, reflexives from several languages have been shown to be exempt from the locality conditions imposed by Condition A, such as Mandarin *ziji*, Icelandic *sig*, and Japanese *zibun*. These exceptions are often known as *exempt* anaphors (Pollard & Sag, 1992).

Two competing approaches have been proposed to capture such exempt anaphors. One approach is the *logophoricity* theory, explaining exempt anaphors by *logophoric* rather than structural constraints (Sells, 1987; Huang & Liu, 2001; Charnavel & Sportiche, 2016). Logophoric constraints are discourse constraints such as mental perspectives that affect referential dependencies between reflexives and their antecedents. The second approach is the long-distance binding (LDB) theory, a pure structural account. Pure structural accounts are those only involve syntactic structural relationships such as c-command/m-command, without considering discourse factors. The LDB theory derives non-local binding via cyclic movement/re-indexing and turns non-local binding into local binding which always obeys Condition A. (e.g., Pica, 1987; Cole et al., 1990; Huang & Tang, 1991; Cole & Wang, 1996). The two accounts make distinct predictions on the distribution of reflexives and the referential dependencies between reflexives and their antecedents.

This paper aims to test the two competing theories by investigating the referential properties of Chinese reflexive *ziji* and *taziji* using an acceptability judgment task. Well-controlled acceptability judgment

tasks may be especially helpful in the study of Chinese anaphors, because the judgments in the literature are not always consistent, which contributes to disagreements in the theories.

On the theoretical side, the results of the present paper shed new light on the debate between logophoricity and LDB theories. On the empirical side, this project not only contributes to providing a more complete picture of the referential properties of *ziji* and *taziji* but also evaluates some of the diagnostics in the syntactic literature for logophoricity contrasts in Mandarin Chinese. In addition, this work has implications on the typology of (non-)local reflexives.

Before introducing the experiments, I provide the theoretical background of (i) logophoricity and LDB theories of *ziji* and (ii) the proposed referential properties of *taziji* in the literature.

3.1.1 Logophoricity and long-distance binding theories of *ziji*

Mandarin *ziji* is categorized as an exempt anaphor, based on observations such as (1). Unlike English *himself*, *ziji* can refer to either the local antecedent *Wangwu* or the long-distance (LD) antecedent *XiaoLi*. To capture binding observations like this, discourse-based logophoricity theories and pure structural LDB theories have been proposed (e.g., Battistella,1989; Pica, 1989; Cole et al., 1990; Cole & Sung, 1994; Cole & Wang, 1996; Tang, 1989; Huang & Tang, 1991; Yu, 1992, 1996; Huang & Liu, 2001).

(1) XiaoLi$_j$ zhidao Wangwu$_k$ xihuan ziji$_{j/k}$

XiaoLi know Wangwu like Refl

'XiaoLi$_j$ knows Wangwu$_k$ likes himself/him$_{j/k.}$'

(Cole et al., 1990)

A. Logophoricity theories and *ziji*

The notion of 'logophor'

The term 'logophor' originally was created to refer to pronominals in African languages referring to the individual 'whose speech, thoughts, feelings, or general state of consciousness are reported' (cf. Hagege, 1974; Clements, 1975). Later this notion was adopted to study exempt reflexives, such as Chinese *ziji* (e.g., Yu, 1992, 1996; Chen, 1992; Huang & Liu, 2001), because exempt anaphors take antecedents that are perspective centers of statements, just like African logophoric pronominals.

The logophoricity theories

From a semantic perspective, Kuno (1972, 1987) claimed that sentences with an embedded pronoun or reflexive referring to the matrix subject could be analyzed as deriving from a direct discourse representation where the pronoun functions as the first-person pronoun. Such a representation is obligatory for sentences with quotative or attitudinal verbs. For instance, (2a) was argued to be transformed from (2b)[18].

(2) a. Ali claimed that he was the best boxer in the world.

b. Ali claimed, "I am the best boxer in the world."

Sells (1987) proposed that the distribution of reflexives was fundamentally determined by syntactic factors and logophoricity had impacts only on non-local reflexives. He further argued that there was no

[18] Previous studies also established a connection between *de se/de* re reading to logophoricity, i.e. logophors can only be interpreted *de se* while anaphoric pronouns are compatible with *de se* or *de re* reading. For details, see Chierchia (1989).

unified notion of logophoricity, but three primitive roles for the antecedent of a logophor: *Source, Self,* and *Pivot,* as specified in (3):

(3) a. Source: the one who is the intentional agent of the communication

b. Self: the one whose mental state or attitude the proposition describes

c. Pivot: the one with respect to whose (space-time) location the content of the proposition is evaluated.

Source is the one who makes the claim - the speaker. Self is the party whose mind is expressed – the attitude holder of the statement. Self may or may not be the speaker. For instance, in 'Bill worried that his girlfriend is leaving him', Bill is not the speaker of this clause, but he is the attitude holder, thus qualified as Self. Pivot, on the other hand, is the one whose point of view in terms of time and space is reported. For example, to describe an event that Mary just flew from London to Boston, John, who is located at Boston, might say 'Mary just came to Boston.' rather than 'Mary just went to Boston.', as the Pivot of this sentence, since John is using his own spacial perspective in the statement.

Ziji as a logophor

As Yu (1992, 1996) pointed out, *ziji* can be sentence-free and refer to the speaker (Source). In (4), there is no available antecedent for *ziji* within the sentences, and *ziji* refers to the Source of the statement. The acceptability of (4) can't be explained by pure structural accounts, because these accounts require the reflexive to be bound by a c-commanding antecedent within the most local tensed clause.

(4) a. Ziji shi xiangxiaren

Refl is peasant

'I am a peasant.'

b. Chule ziji, zhiyou sangeren zancheng.

Except for Refl, only three people approve

'Besides myself, only three people agree.'

(Yu, 1992)

B. Long-distance binding (LDB) theories of *ziji*

This section provides an overview of the three major types of structural accounts on long-distance *ziji*: (a) cyclical re-indexing; (b) head movement; and (c) IP adjunction. Under all three accounts, long-distance binding should be converted to local binding via a syntactic operation, such as cyclical re-indexing (Tang, 1989), head-movement (e.g., Battistella, 1989; Pica,1989; Cole et al., 1990; Cole & Sung, 1994; Cole & Wang, 1996), or IP adjunction (Huang & Tang, 1991). In this way, *ziji* can be locally bound by its antecedent at an abstract level (e.g., Logical Form), thus obeying Condition A.

In addition to explaining the exempt binding conditions of *ziji*, these accounts also attempt to derive other relevant properties of *ziji*, such as *ziji* being mono-morphemic or only referring to the subject (subject-orientation) via the re-indexing/movement mechanism (e.g., Battistella, 1989; Pica, 1989; Cole et al., 1990; Cole & Sung, 1994; Cole & Wang, 1996; Tang, 1989). Historically, an important criterion in comparing and evaluating various LDB theories has been whether the theories can capture these properties. However, it turns out that *ziji* is neither mono-morphemic nor subject-oriented (see Experiment 1), thus undermining the original motivations for these proposals.

Cyclical re-indexing

Tang (1989) attempted to capture the long-distance (LD) binding of *ziji* via an optional feature-copying rule and a cyclic re-indexing rule. In this approach, *ziji* is in fact *pro-ziji* where *pro* transfers its phi-features to *ziji*. The phi-features can't be changed after being fixed, and therefore *ziji* can't be bound by a LD antecedent whose phi-features are not compatible with the local antecedent. This account derives the non-local binding of *ziji*. The feature-copying and re-indexing rules only apply to *ziji*, but not *taziji*, because *taziji* already has fixed phi-features and it is assigned an unchangeable referential feature on the first local re-indexing cycle. Thus, Tang (1989) predicted *taziji* and reflexives with fixed phi-features in other languages are purely local.

However, Tang (1989)'s prediction is not borne out. As will be shown in Experiment 3, Chinese *taziji* is not a local anaphor. Furthermore, there are counterexamples from other languages: in Italian, both long-distance (*se* and *proprio*) and local (*se stesso*) reflexives are lexically specified with person features (Charnavel et al., 2017).

Head movement

Head movement accounts of long-distance binding (e.g., Battistella, 1989; Pica,1989; Cole et al., 1990; Cole & Sung, 1994; Cole & Wang, 1996) typically assume that *ziji* is a mono-morphemic head and it moves to I^0 of the most local IP and optionally moves I^0-to- I^0 to a higher IP at Logical Form (LF). In this way, *ziji* can move to the matrix INFL position and be locally bound by the matrix subject, which is in line with Condition A.

On the surface, there appear to be several benefits of the head movement approach, but, these advantages disappear under scrutiny. The first purported advantage is that these accounts can derive the

referential properties of reflexives via their morphologies. For instance, the head movement accounts proposes that only mono-morphemic reflexives are qualified as heads to undergo head movement. Based on the claim that *ziji* is mono-morphemic and exempt, while *taziji* is bi-morphemic and local, the head movement account provides a natural explanation for this distinction between these two reflexives. That is, *ziji* is non-local, because it is mono-morphemic and can move at LF to co-index with the LD antecedent, while *taziji* cannot, since it is not mono-morphemic and not a head. However, contrary to assumption, it turns out that *ziji* is in fact bi-morphemic (Liu, 2016; Wong, 2017; Reuland, 2018). Furthermore, *taziji* can also be exempt (see Experiments 2&3 in this paper). Thus it is difficult to conclude a correlation between reflexives' referential properties and their morphological features. The second purported advantage of the head movement account is based on another invalid assumption about *ziji*: the so-called subject orientation. It was proposed that *ziji* can only refer to subjects that c-command it (e.g., Tang, 1989), which can be derived via the I^0-to- I^0 movement, as the landing position of *ziji* is at INFL and only the subject c-commands INFL. However, results from the literature and Experiment 1 here show that *ziji* can naturally refer to a non-subject and non-c-commanding antecedent (Xu, 1993, 1994; Yu 1992, 1996).

IP adjunction

Another movement account is the IP adjunction analysis. Huang & Tang (1991) noted that LD binding for *ziji* could be established across an island (such as relative clause in 5), and head movement should obey locality conditions, which may cast doubt on the head movement account. Hence, Huang & Tang (1991) proposed the IP adjunction account in which *ziji* adjoined to IP in a successive cyclic fashion (A'-movement). More specifically, the LD reflexive adjoins in LF to an IP, and from the IP-adjoined position, it is coindexed with its antecedent. Based on the hypothesis that Subjacency and Condition on

Extraction Domains (CED) do not obtain in LF, this account predicts LD *ziji* does not exhibit Subjacency or CED effects.

(5) Zhangsan$_k$ bu xihuan [$_{NP}$[$_{CP}$ neixie piping **ziji**$_k$] de ren]
Zhangsan not like those criticize Refl DE people
'Zhangsan$_k$ does not like those people who criticized him$_k$'

3.1.2 Tests distinguishing the logophoricity and LDB theories

One way to disentangle the two kinds of competing theories is to test whether *ziji* can take a non c-commanding antecedent while manipulating the logophoric status of the antecedent, as in (6). In (6a), 'According to *XiaoLi*' expresses *XiaoLi*'s attitude, while 'speaking of *XiaoLi*' in (b) usually expresses the speaker's rather than *XiaoLi*'s perspective (cf. Kuno, 1987; Sells, 1987). The logophoricity theory thus predicts (6a) to be more acceptable than (b). The LDB theory predicts both (6a) and (b) to be ungrammatical, because the antecedents do not c-command *ziji.*

(6) a. Ju XiaoLi$_k$ shuo, zhejianshi shanghai-le ziji$_k$
According to XiaoLi say, this event hurt-Perf Refl
'**According to** XiaoLi$_k$, this event hurt himself$_k$.'

b. Shuodao XiaoLi$_k$, zhejianshi shanghai-le ziji$_k$
Speaking of XiaoLi , this event hurt-Perf Refl
'**Speaking of** XiaoLi$_k$, this event hurt himself$_k$.'

Another way is to test whether *ziji* can always take a c-commanding antecedent both in logophoric and non-logophoric conditions, as in the Adjunct Clause structures in (7):

(7) a. XiaoLi$_k$ likai-le gongsi, yinwei nvjingli piping-le ziji$_k$

XiaoLi leave-Pef company, because female manager criticize-Pef Refl

'XiaoLi$_k$ left the company, **because** the female manager criticized himself$_k$.'

b. yinwei nvjingli piping-le ziji$_k$, suoyi XiaoLi$_k$ likai-le gongsi.

because female manager criticize-Pef Refl, thus XiaoLi leave-Pef company

'Because the female manager criticized himself$_k$, XiaoLi$_k$ thus left the company.'

c. Dang nvjingli piping-le ziji$_k$ de shihou, XiaoLi$_k$ likai-le gongsi.

When female manager criticize-Pef Refl DE time, XiaoLi leave-Pef company

'When the female manager criticized himself$_k$, XiaoLi$_k$ left the company.'

A 'because'-clause denotes the mental status of the main-clause subject *XiaoLi*, and can occur either after (6a) or before the main clause (7b). A 'when'-clause (7c) can only precede the main clause, and is from the perspective of the speaker, not *XiaoLi* (Huang & Liu 2001). In (7abc), the antecedent c-commands *ziji*[19], which means *XiaoLi* is a legitimate antecedent for *ziji* in all 3 sentences according to the LDB theory. The logophoricity theory thus predicts (7ab) to be more acceptable than (7c) and, but the LDB theory predicts no acceptability difference among (7abc).

[19] A possible independent test for the c-commanding relationship between *XiaoLi* and *ziji* is that the bound variable meaning is available for pronouns in *because* and *when*-clauses in Chinese. For instance, 'When the manager criticized his child, every parent was unhappy.' can have the interpretation that every parent is unhappy for his own child, not necessarily for a certain guy's child. The same holds for *because*-clause too.

3.1.3 Referential properties of *taziji* in the literature

The mostly widely accepted view of *taziji* is that it is strictly local, not violating Condition A (e.g., Tang,1989; Huang & Tang, 1991; Cole & Sung, 1994; Cole et al., 2005).

A. *Taziji* is a local anaphor

The proposal that *taziji* is a local anaphor is based on contrasts such as between (8a) and (8b). In (8a), *ziji* can refer to either local *Lisi* or LD *Zhangsan*, but *taziji* can only refer to the local antecedent *Lisi* in (8b).

(8) a. Zhangsan$_i$ renwei Lisi$_j$ hai-le ziji$_{i/j}$

'Zhangsan$_i$ thought that Lisi$_j$ hurt himself$_{i/j}$'

b. Zhangsan$_i$ renwei Lisi$_j$ hai-le taziji$_{*i/j}$

'Zhangsan$_i$ thought that Lisi$_j$ hurt himself$_{*i/j}$'

(Huang & Tang, 1991)

Under this view, the only scenario in which *taziji* can be exempt from Condition A is sub-command (Tang, 1989; Dillon et al., 2016): just like *ziji*, *taziji* can be bound by an animate NP within an inanimate subject, as in (9). Note that there is gender marking (male and female for animal NPs; 'neutral' for inanimate NPs) on the 1st morpheme *ta-* in the writing system, though the pronunciation of *taziji* for different gender features is the same.

(9) Zhangsan$_j$ de jiaoao hai-le ziji$_j$/taziji$_j$ (M)

Zhangsan DE pride destroy-Pef Refl.M

'Zhangsan$_j$'s pride destroyed himself$_j$.'

(Tang, 1989; Dillon et al., 2016)

B. *Taziji* can be exempt

The earliest articles arguing that *taziji* is exempt are Yu (1992,1996) and Pan (1997). Yu (1992) pointed out that the reflexive *taziji* could be bound across an animate local subject and therefore be exempt. In (10), for example, *taziji* cannot be bound by the local subject *Mali* due to mismatch in gender, but it can refer to the matrix subject *Yuehan*.

(10) Yuehan$_i$ jiao Mali$_j$ chuipeng taziji$_{i/*j}$ (Male).

'John$_i$ asked Mary$_j$ to flatter himself$_{i/*j}$'

(Yu, 1992)

This idea was later developed in Yu (1996), which further stated that besides being sub-commanded in (9), *taziji* also shares other properties with *ziji*, such as being LD bound and sentence-free. Yu (1996) follows Reinhart and Reuland (1993) and defines the local domain as the co-argumenthood domain. In (11), *taziji* can be co-indexed with LD *Zhangsan* across the embedded subject 'someone'. In (12), *taziji* is not bound within the clause; instead, it refers to a person accessible in the context, such as a person mentioned earlier in this conversation.

(11) Zhangsan$_j$ zhidao youren zai genzong taziji$_j$ (M)

Zhangsan know someone Progressive follow Refl.M

'Zhangsan$_j$ knows someone is following him$_j$.'

(Yu, 1992)

(12) Ni wen taziji (M)

You ask Refl.M

'You ask himself'

(Yu, 1996)

Though Yu (1996) regards exempt *taziji* as logophoric, no minimal pair contrasts with manipulation of logophoricity was provided, so it's unclear what kind of logophoric constraints *taziji* should obey.

Similar to Yu (1992, 1996), Pan (1997) pointed out that *taziji* can be non-local. According to Pan (1997), *taziji* can be bound across an inanimate subject, such as in (13). *Taziji* can also skip a local human subject (a *pro*) in the relative clause and refer to the matrix subject, as in (14).

(13) Zhangsan$_k$ shuo nabenshu hai-le taziji$_k$ (M)

Zhangsan say that book hurt-Pef Refl-M

'Zhangsan$_k$ said that book hurt him$_k$.'

(Pan, 1997)

(14) John$_i$ meiyou zhaodao yifeng [S pro$_j$ xie gei (ta)-ziji$_{i/*j}$] de xin.

John not-have find-out one write to he-self DE letter

'John$_i$ has not found one letter which is written to him$_i$.'

(Pan, 1997)

Unlike Yu (1992, 1996), Pan (1997) claims it is very hard to have an object *taziji* bound by an NP across its local human subject, but it is possible to construct contexts to facilitate the use of an object *taziji.* The matrix subject in (15) is the preferred antecedent for *taziji,* as it is pragmatically odd to have *Bill* as the antecedent of *taziji*.

(15) John$_i$ zhidao Bill$_j$ bu xihuan taziji$_{i/?j}$, suoyi pro$_i$ meiyou qu qiu ta$_{j.}$

John know Bill not like Refl-M, so he not to beg him.

'John$_i$ knows Bill$_j$ does not like him$_i$, so he$_i$ didn't beg him$_j$. '

In addition, Pan (1997) claims that reflexive *taziji* is contrastive if an animate subject is skipped. Though reflexive *taziji* is contrastive in many circumstances, there exist three exceptional cases in which *taziji* is not necessarily contrastive as in (16).[20]

(16) a. There is no intervening subject;

b. There is only an intervening inanimate subject;

c. There is only a sub-commanding antecedent.

In sum, there are two primary distinct hypotheses about *taziji*. One is that it is a strictly local reflexive, the other one is that *taziji* can be exempt. All the data in (10)-(15) from Yu and Pan's work seem to challenge the traditional proposal of applying Condition A directly to *taziji*.

[20] Pan (1998) provides an independent diagnostic for contrastive and non-contrastive *taziji*. These two elements are phonologically different: *ziji* is stressed in contrastive *taziji* but not in noncontrastive *taziji*. There are two potential issues with this diagnostic. First, empirical phonologically evidence needs to be provided for such a diagnostic. As a native speaker of Mandarin, I'm personally not sure if (16) is valid. Second, it is unclear how and why syntactic/morphological properties of *taziji* interact with its pronunciation.

The remainder of this paper is structured as follows. Experiment 1 offers an experimental investigation on non-local co-indexing of *ziji* to tease apart the two competing theories. The results show that *ziji* is sensitive to logophoricity effects and can refer to a non-subject and non-c-commanding antecedent. Experiment 2 tests whether *taziji* can be exempt from Condition A. Our results reveal that *taziji*, though usually considered as a local anaphor (e.g., Huang & Tang, 1991; Cole et al., 2006 i.a.), can in fact similarly be non-local bound as *ziji*. Experiment 3 further investigates whether non-local *taziji* can be better explained by logophoricity or LDB theories. The results yield that contrary to the LDB theories, *taziji* can be exempt from binding under logophoric conditions.

3.2 Experiment 1: Disentangling Logophoricity and LDB Theories with *ziji*

Experiment 1 aimed to tease apart the logophoricity and LDB theories via an acceptability judgment task. The logophoricity theory and LDB theory make distinct predictions about the syntactic distribution of reflexives and the referential dependencies between reflexives and their antecedents. The LDB theory requires reflexives be c-commanded by their antecedents. The logophoricity theory, however, predicts that reflexives need not be c-commanded by their antecedents if they are logophoric.

Thus, the LDB theory predicts all possible antecedents of *ziji* should c-command it, and a non-c-commanding antecedent is not acceptable for *ziji*. The logophoricity theories predict that *ziji* can refer to a non-c-commanding, logophoric antecedent.

3.2.1 Participants

80 Mandarin speakers participated in this experiment via a crowdsourcing platform, Witmart, in exchange for $2.

3.2.2 Materials and Design

The acceptability task includes two distinct structures involving contrasts in logophoricity, as in Section 1.2, copied as below (17) - (18).

(17) Prepositional Phrase structure

a. Ju XiaoLi$_k$ shuo, zhejianshi shanghai-le ziji$_k$ [ziji = XiaoLi]

According to XiaoLi say, this event hurt-Perf Refl

'**According to** XiaoLi$_k$, this event hurt himself$_k$.'

b. Shuodao XiaoLi$_k$, zhejianshi shanghai-le ziji$_k$

Speaking of XiaoLi , this event hurt-Perf Refl

'**Speaking of** XiaoLi$_k$, this event hurt himself$_k$.'

(18) Adjunct Clause structure

a. XiaoLi$_k$ likai-le gongsi, yinwei nvjingli piping-le ziji

XiaoLi leave-Pef company, because female manager criticize-Pef Refl

'XiaoLi$_k$ left the company, **because** the female manager criticized himself$_k$.'

b. yinwei nvjingli piping-le ziji$_k$, suoyi XiaoLi$_k$ likai-le gongsi.

because female manager criticize-Pef Refl, thus XiaoLi leave-Pef company

'Because the female manager criticized himself$_k$, XiaoLi$_k$ thus left the company.'

c. Dang nvjingli piping-le ziji$_k$ de shihou, XiaoLi$_k$ likai-le gongsi.

When female manager criticize-Pef Refl DE time, XiaoLi leave-Pef company

'When the female manager criticized himself$_k$, XiaoLi$_k$ left the company.'

17 pairs of sentences were constructed for the contrast between 'speaking of' and 'according to' in (17). 7 pairs of clauses were made for the 'when'/'because'-clause contrast in (18). Though *ziji* in (17) is in principle not ambiguous in terms of reference, to avoid potential misunderstanding in (18) and keep all the test stimuli with the same format, the intended reference of *ziji* was indicated in brackets [ ] as in (17a) for all sentences.

To check if participants paid attention to the task, 10 attention check sentences were also included, which consisted in 5 uncontroversially acceptable and 5 uncontroversially unacceptable ones, as in (20). Only one member of each item in test stimuli was presented to each person, so that each participant saw 34 sentences in total, which were presented in random order. Participants were asked to provide judgments on a binary scale ('acceptable' or 'unacceptable').

(20) a. Sample acceptable sentence

Zhaogang yudao-le Wangyan, ta(M) juede zhegenvren henyoumeili [zhegenvren=Wangyan]

Zhaogang meet-Pef Wangyan, he find this woman charming

'Zhaogang met Wangyan, and he finds this woman charming.' [this woman=Wangyan]

b. Sample unacceptable sentence

Xiayudeshihou, Xiaohong kandao Wangchen qiang-le nagenanren de san.[nagenanren=Wangchen]

When it rains, Xiaohong observe Wangchen rob-Pef that man DE umbrella

'When it rains, Xiaohong observes that Wangchen robbed that man's umbrella.'

[that man=Wangchen]

3.2.3 Results

Data from participants who didn't achieve at least 80% accuracy on the attention check materials were excluded. Responses from the remaining 71 participants were analyzed. Mixed-effects logistic regressions in *lme4* package in *R* were applied to acceptability judgements of the three tested structures.

In all the experiments in this paper, we report models with the maximal random effect structures that were appropriate for the experimental design and converged (Barr et al., 2013). The random effect structures were chosen by model comparison based on likelihood-ratio test and/or Bayesian Information Criterion (anova() or BIC() function in R)[21].

For the PP structures, a mixed-effects logistic regression was applied. We entered *logophoric condition* as the predictor, with random intercepts for *subject* and *item*, as well as by-subject and by-item *logophoric condition* slopes. The model revealed that *ziji* with logophoric 'according to' was significantly more acceptable than with non-logophoric 'speaking of' (β=-5.43, Z=-5.77, $p<0.001$), as shown in Figure 1. These results were more in line with the logophoricity theory than the LDB theory.

In addition, the mean proportion of 'acceptable' response for the 'according to' case was 0.97 (very close to ceiling 1), suggesting *ziji* can naturally refer to a non-c-commanding and non-subject antecedent. Interestingly, in the 'speaking of' condition, the antecedent was neither logophoric, nor c-

[21] We first fit the model with the maximal random effect structure. We reported results of this model, if it converged. Otherwise, we started from the random intercepts-only model, including random intercepts for subject and item, and performed forward model comparison using likelihood-ratio tests (anova function in R). Random slopes will be included in the final model, if model fitness is significantly improved. In cases where we needed to choose between two non-nested random effect structures (e.g., (1 + logophoric condition | subject) + (1 | item) vs. (1 | subject) + (1+ logophoric condition | item)) and likelihood-ratio test cannot be deployed, we reported the model with a lower BIC.

commanded the reflexive, but around 50% responses judged the sentence as ‘acceptable’, which suggests that logophoricity may be gradient rather than dichotomous.

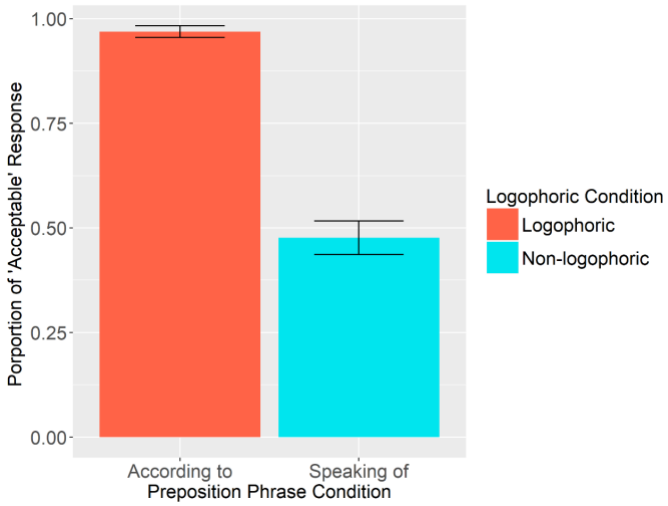


Figure 1. Proportion of ‘acceptable’ response in Prepositional Phrase structures by logophoric condition (logophoric vs. non-logophoric). The error bars show 95% confidence interval.

For Adjunct Clause structures, *logophoric condition* was entered as the predictor. The model also included random intercepts for *participant* and *item*, as well as by- participant *logophoric condition* slope. We dummy coded the three levels in *logophoric condition*, with the non-logophoric ‘when’-clause as the reference level and comparing the two ‘because’-clause levels to it. As predicted by the logophoricity theory, *ziji* in ‘because’-clauses before or after the main clause were significantly more acceptable than in ‘when’-clauses ($\beta s > 4.8$, $Zs > 3.07$, $ps < 0.01$), since antecedents are more mentally involved in sentences denoting causal relationships, as plotted in Figure 2. In contrast, the LDB theory does not explain the significant difference obtained between distinct adjunct clauses containing *ziji*, since antecedents in main clauses c-command *ziji* in all the sentences.

We again found that around 50% responses judged *ziji* in non-logophoric ‘when’-clauses as 'acceptable’, which suggests that although the logophoricity account may be on the right track, it does not provide a comprehensive explanation for the observed data.

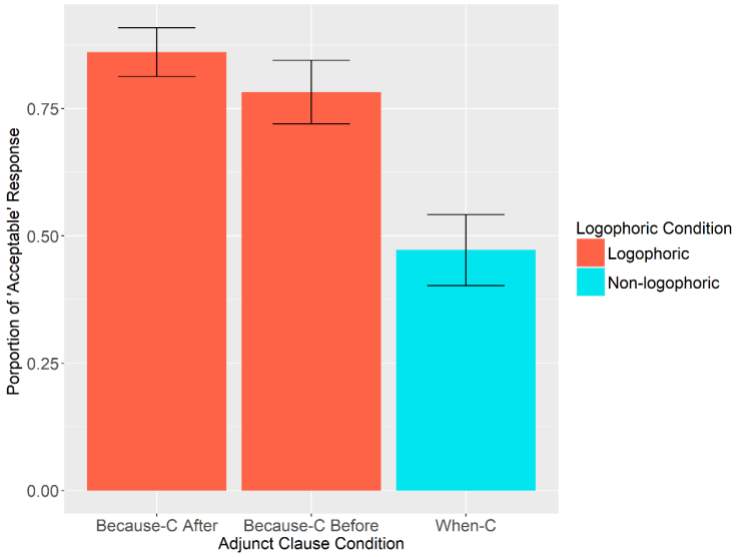


Fig. 2. Proportion of ‘acceptable’ response in the Adjunct Clause contrast by logophoric condition (logophoric vs. non-logophoric). The error bars show 95% confidence interval.

3.2.4 Discussion

The two tests in experiment 1 thus support the hypothesis that non-local *ziji* is a logophor, not a LD anaphor, given that (i) *ziji* can naturally refer to a logophoric but non-c-commanding antecedent, and (ii) *ziji*, when c-commanded by the antecedent, has significant acceptability contrast between logophoric ‘because’-clauses and non-logophoric ‘when’-clauses.

In addition, the subject orientation property of *ziji* in previous literature might be an artifact of its logophor property, given that the subject is usually the perspective taker of the whole clause. The PP

condition has shown that *ziji* can be bound by a non-c-commanding antecedent, suggesting it is not the structural feature of the subject such as c-command that makes the subject accessible to *ziji*.[22]

Furthermore, given that non-logophoric *ziji* is also rated as fairly natural (around 50% 'acceptable' responses), a more refined theory of logophoricity is needed for a comprehensive explanation for these observations.

As for the typology of reflexives and their binding conditions, the traditional claim is that long-distance reflexives are usually monomorphemic, while local reflexives tend to consist of more than one morpheme (Giorgi, 1984; Pica, 1987). However, it seems that not all monomorphemic reflexives are LD (e.g., German *sich*), and there is no clear correlation between being morphologically complex and being local universally. Both Chinese *ziji* (and its equivalents in Cantonese, Min, and Dong, a minority language in China) and Japanese *zibun* are exempt and complex (Kishida, 2012; Liu, 2016). The first morpheme *zi-* in Chinese *ziji* (21a) and Japanese *zibun* (21b) could independently form words with other morphemes. In addition, the second morpheme *ji* in Chinese *ziji* can also construct words independently (22).

(21) a. Chinese *zi-*:

zi-kua

self-brag

b. Japanese *zi-*:

zi-ritu

[22] Besides the c-command account on subject orientation, another previous approach claims that subject orientation can be attributed to thematic hierarchy. For counter examples of the latter account, see Charnavel, Huang, Cole and Hermon (2017).

self-establish

(22) lv-**ji**

discipline-self

Besides the pure structural accounts, several mixed approaches have also been proposed, stating that *ziji* is syntactically bound in the local domain and logophorically bound in the non-local domain, such as Xue et al (1994) and Huang & Liu (2001), though the definitions of local domain differ. This paper mainly tests non-local *ziji*, and I leave the issue whether *ziji* is plain or logophoric in the local domain for future research. There have been studies showing that local reflexives might be logophoric too (Sloggett & Dillon, 2018 on English *himself*).

3.2.5 Summary of Experiment 1

Tables 1 and 2 offer a summary of the contributions made by experiment 1 to current syntactic theories on *ziji* binding and to the syntactic diagnostics used in these studies.

Table 1: Implications to the theoretical claims about reflexive *ziji*

Theoretical issue	**Previous studies**	**Present study**
Explanation of non-local binding	Debate between long-distance binding and logophoricity theories	Logophoricity theory, not pure structural accounts
C-commanding	Required (pure structural accounts)	Not necessary
Subject orientation	Yes (pure structural accounts)	No, might be artifact of logophoricity effects
Morphology	Mono-morphemic, simplex (head-movement accounts)	Bi-morphemic, complex

Table 2: Empirical implications on diagnostics of logophoricity in Mandarin Chinese

Diagnostics	**Test results**
Speaking of vs. According to	Works as in the literature
When-clause vs. Because-clause	Works as in the literature

3.3 Experiment 2: *Taziji* - Local or Exempt?

As discussed above, it is still debated whether *taziji* is exempt. Consequently, Experiment 2 aimed to test whether *taziji* can be LD bound across a local animate subject with distinct types of embedded verbs: 'mutual-direction' verbs and 'other-direction' verbs. The former type of verbs represents actions that one can do to oneself or others, such as *zeguai* ('blame'); the latter type of verbs denote actions that one can only do to others like *genzong* ('follow'). A norming study was performed to diagnose the two groups of verbs.

3.3.1 Participants

42 Mandarin speakers participated in the norming study and another group of 60 participants did the acceptability judgment task for *taziji* via a crowdsourcing platform, Witmart, in exchange for $2.

3.3.2 Norming Study

The goal of this norming study was to distinguish between mutual and other-direction verbs. Given that *ziji* can refer to either the local or the LD subject, especially in sentences with the matrix verb 'say', we constructed test items where both the local *Lijun* and LD *Wanggang* are legitimate antecedents for *ziji* and participants were asked to choose their preferred antecedent between these two options, as in (23). The matrix and embedded subjects were common Chinese proper names with the same gender (stereotypically) and the matrix verb is 'say' in all the sentences. 100 embedded verbs were tested.

(23) **Sentence**: Wanggang shuo Lijun xihuan ziji.

'Wanggang says Lijun likes *ziji*.'

Question: Who does *ziji* refer to in this sentence?

A. Wanggang

B. Lijun

The more participants choose the LD antecedent, the more 'other-direction' the embedded verb is. If there is a similar number of responses choosing the local and the LD subjects for a certain verb, then it will be regarded as a 'mutual-direction' verb. Our threshold for 'other-direction' verb is over 75% responses choosing the LD subjects, and 25-75% responses choosing the LD subjects for 'mutual-direction' verbs. Among the 100 tested embedded verbs, 48 verbs were collected including 24 'other-' and 24 'mutual-' direction verbs.

3.3.3 Materials and Design

For each of the 48 verbs, a set of sentences of 3 conditions were constructed, the local, LD and no match condition. A set of sample stimuli for 'mutual-direction' and 'other-direction' verbs is illustrated in (24) and (25) respectively. We manipulated the gender features of the local and LD subjects and *taziji.* There is gender marking on the 1st morpheme of *taziji* in the writing system (他自己 for 'himself' and 她自己 for 'herself'). In the local match case, the gender feature of *taziji* is in line with the local antecedent (24a). In the LD case, *taziji*'s binder matches the LD antecedent (24b), while the no match case includes local and LD antecedents whose gender features do not match *taziji* (24c).

(24) 'Mutual-direction'

a.Local match *Zhangxiansheng*$_{i}$ *shuo Lixiaojie*$_{k}$ *piping-le taziji*$_{*i/k}$ (F).她自己
'Mr.Zhang$_{i}$ said Ms.Li$_{k}$ criticized herself$_{*i/k}$.'

b.Long-distance match *Zhangtaitai*$_{i}$ *shuo Lixiansheng*$_{k}$ *piping le taziji*$_{i/*k}$ (F).她自己
'Ms.Zhang$_{i}$ said Mr.Li$_{k}$ criticized herself$_{i/*k}$.'

c. No match/ Ungrammatical control *Zhangxiansheng*$_{i}$ *shuo Lishushu*$_{k}$ *piping-le taziji*$_{*i/*k}$ (F).她自己
'Mr.Zhang$_{i}$ said Uncle Li$_{k}$ criticized herself$_{*i/*k}$.'

(25) 'Other-direction'

a.Local match *Zhangxiansheng*$_{i}$ *shuo Lixiaojie*$_{k}$ *genzong-le taziji*$_{*i/k}$ (F).她自己
'Mr.Zhang$_{i}$ said Ms.Li$_{k}$ followed herself$_{*i/k}$.'

b.Long-distance match *Zhangtaitai*$_{i}$ *shuo Lixiansheng*$_{k}$ *genzong-le taziji*$_{i/*k}$ (F).她自己
'Ms.Zhang$_{i}$ said Mr.Li$_{k}$ followed herself$_{i/*k}$.'

c. No match/ Ungrammatical control *Zhangxiansheng*$_{i}$ *shuo Lishushu*$_{k}$ *genzong-le taziji*$_{*i/*k}$ (F).她自己
'Mr.Zhang$_{i}$ said Uncle Li$_{k}$ followed herself$_{*i/*k}$.'

If *taziji* is a local anaphor, LD match (24&25b) should be as unacceptable as the no match control (24&25c). If *taziji* can be exempt, then the LD match case should be at least better than the no match condition with mutual-direction verbs in (24) and might also be better than the local match condition with other-direction verbs in (25), due to pragmatic or semantic factors.

48 sets of sentences were constructed for the 48 embedded verbs with 3 conditions in each set. To check if participants have paid attention to the task, 10 attention check sentences were also included, with 5 acceptable and 5 unacceptable ones, as in Experiment 1. Only one member of each item in the 48 test stimuli was presented to each person, so each participant saw 58 sentences, with the order randomized for each participant. Participants were asked to rate sentences on a 1 (very unacceptable)--7 (very acceptable) Likert Scale.

3.3.4 Results

Data from participants who didn't judge the attention checks with at least 80% accuracy were excluded, leaving responses from 57 participants. Two mixed-effects linear regressions were fit for the two groups of verbs separately.

In the mutual-direction condition, *match condition* was entered as the predictor. The random effect structures included random intercepts for *participant* and *item* and by-participant *match condition* slope. The three levels of *match condition* (LD match, local match, no match) were dummy coded, with LD match as the reference level. The ratings of LD match *taziji* were significantly higher than no match/ungrammatical control (β=-2.42, t=-8.84, p<0.001), though lower than the local match (β=0.64, t=2.95, p<0.01), as Fig. 4. Contrary to the widely accepted view of *taziji* strictly obeying Condition A, the results indicated that *taziji* can refer to a LD antecedent.

For the other-direction verbs, we included *match condition* as the predictor and random intercepts for *participant* and *item* in the model. The same way of dummy coding was deployed. Consistent with the exempt *taziji* hypothesis, LD *taziji* was rated significantly more acceptable than both local *taziji* and no match *taziji* (βs < -2.59, ts < -10, ps<0.001), as shown in Figure 4.

Results of both groups of verbs indicated that *taziji* can naturally refer to a non-local antecedent, thus being exempt from Condition A. In fact, LD binding of *taziji* can be even more acceptable than local binding under some circumstances, such as with the other-direction verbs.

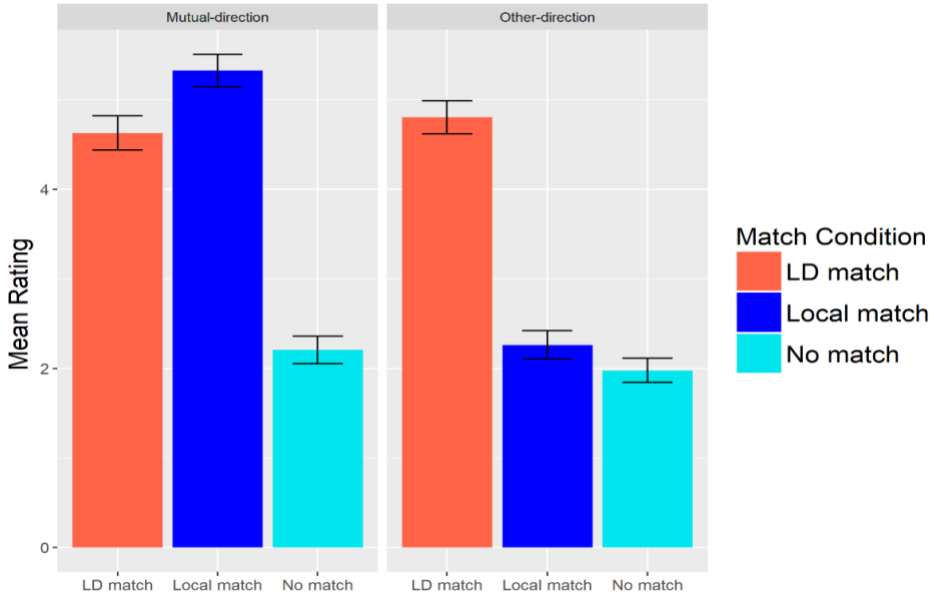


Fig.4. The mean ratings by verb direction ('mutual' vs. 'other') and match condition (local vs. LD vs. no match). Error bars show 95% confidence interval.

3.3.5 Discussion

Overall, the results suggested that *taziji* can in fact be LD bound: it is not a local anaphor and its binding conditions are sensitive to verb directions.

The results of Experiment 2 showed LD *taziji* was significantly more acceptable than the ungrammatical control but less acceptable compared to local *taziji*, which might be the reason why some previous studies regarded *taziji* as a local anaphor.

The fact that *taziji* can be exempt from Condition A again demonstrates that morphologically complex reflexives can be non-local. There is no clear correlation between being morphologically complex and being local.

3.4 Experiment 3: A Logophoric Explanation for *taziji*

The results of Experiment 2 show that *taziji* can be exempt from Condition A, similar to *ziji*. A theory that can explain exempt *taziji* is needed. We can start with checking whether exempt *taziji* is a logophor or a LD anaphor. Though most logophoricity diagnostics in Chinese were proposed for *ziji*, these in principle can also be applied to *taziji*, since these tests mainly target whether the potential antecedents are logophoric or not.

The goal of Experiment 3 is to test if *taziji* is also sensitive to logophoric conditions and how similar/different exempt *ziji* and *taziji* are. If non-local *ziji* and *taziji* are similarly affected by logophoricity, then a unified theory is potentially possible for the two Chinese reflexives.

3.4.1 Participants

80 Mandarin speakers participated in this experiment via a crowdsourcing platform, Witmart, in exchange for $2.

3.4.2 Materials and Design

The design and materials are very similar to Experiment 1. Sentences in (26)-(27) correspond to Prepositional Phrase and Adjunct Clause structures, respectively.

The logophoricity theory predicts (26a) to be more natural than (26b), while the LDB theory predicts both (26a) and (26b) to be ungrammatical. The LDB theory makes the prediction that there is no acceptability contrast between (27ab) and (27c), while the logophoricity theory predicates (27ab) better than (27c).

(26) a. Ju Xiaowang$_k$ shuo, zhejianshi shanghai-le taziji$_k$ *[taziji = Xiaowang]*

According to Xiaowang say, this event hurt-Perf Refl

'**According to** Xiaowang$_k$, this event hurt himself$_k$.'

b. Shuodao Xiaowang$_k$, zhejianshi shanghai-le taziji$_k$

Speaking of Xiaowang , this event hurt-Perf Refl

'**Speaking of** Xiaowang$_k$, this event hurt himself$_k$.'

(27) a. Xiaowang$_k$ likai-le gongsi, yinwei nvjingli piping-le taziji$_k$

Xiaowang leave-Perf company, because female manager criticize-Perf Refl

'Xiaowang$_k$ left the company, **because** the female manager criticized himself$_k$.'

b. yinwei nvjingli piping-le taziji$_k$, suoyi Xiaowang$_k$ likai-le gongsi.

becausefemale manager criticize-Perf Refl, thus Xiaowang leave-Perf company

'Because the female manager criticized himself$_k$, Xiaowang$_k$ thus left the company.'

c. Dang nvjingli piping-le taziji$_k$ de shihou, Xiaowang$_k$ likai-le gongsi.

When female manager criticize-Pef Refl DE time, Xiaowang leave-Perf company

'When the female manager criticized himself$_k$, Xiaowang$_k$ left the company.'

17 pairs of sentences were constructed for the contrast between 'speaking of' and 'according to' in (26). 7 pairs of clauses were made for the 'when'/'because'-clause contrast in (27). Though *taziji* in those sentences are in principle not ambiguous in terms of reference (given that it is marked in gender), to be consistent with Experiment 1, the intended reference of *taziji* was indicated in brackets [ ] for all

sentences**,** as in (26a). Ten attention check sentences were combined with test stimuli to ensure that participants paid attention to the task, including 5 acceptable and 5 unacceptable ones, similar to Experiment 1. Only one member of each item in the test stimuli was presented to each person, so each participant saw 34 sentences and the order was randomized. Participants were asked to provide judgments on a binary scale ('acceptable' or 'unacceptable').

3.4.3 Results

Data from subjects who didn't judge the attention checks with at least 80% accuracy were excluded. Responses from the remaining 78 participants were analyzed. Three mixed-effects logistic regressions in *lme4* package in *R* were applied to data of the three structures separately.

For the PP structures, *logophoric condition* was entered as the predictor. Random intercepts for *participant* and *item* as well as by-participant *logophoric condition* slope were included as the random intercepts. *Taziji* with logophoric 'according to' was significantly more acceptable than with non-logophoric 'speaking of' (β=-1.67, Z=-2.37, $p<0.02$). These results (i) provided further evidence that *taziji* does not obey Condition A, as it can naturally refer to a non-c-commanding and non-subject antecedent; and (ii) suggested that similar to *ziji*, *taziji* is also sensitive to logophoricity contrasts such as perspective taking.

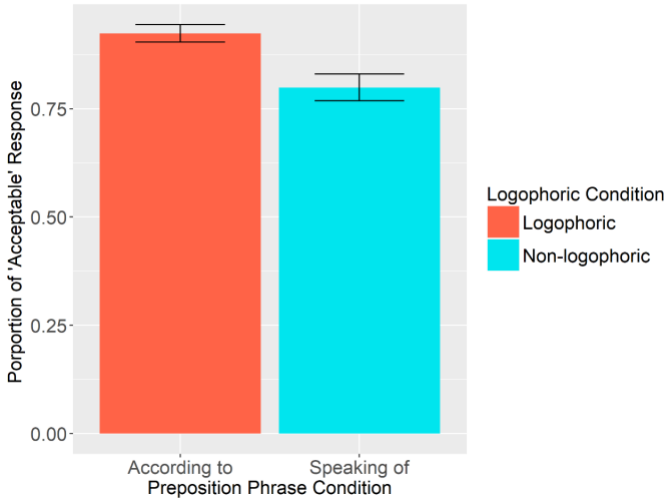


Fig.5. Proportion of 'acceptable' response in Preposition Phrase contrast by logophoric condition (logophoric vs. non-logophoric). The error bars show 95% confidence interval.

For the Adjunct Clause structures, we entered *logophoric condition* as the predictor and random intercepts for *particpant* and *item*. We dummy coded the three levels in *logophoric condition*, with the non-logophoric 'when'-clause as the reference level and comparing the two 'because'-clause levels to it. Similar to *ziji*, *taziji* in 'because'-clauses before or after the main clause were significantly more acceptable than in 'when'-clauses ($\beta s > 0.95$, $Zs > 2.74$, $ps < 0.01$), as antecedents are more mentally involved in sentences denoting causal relationships, as in Fig.6. These results provided another piece of supportive evidence for *taziji* being exempt and logophoric.

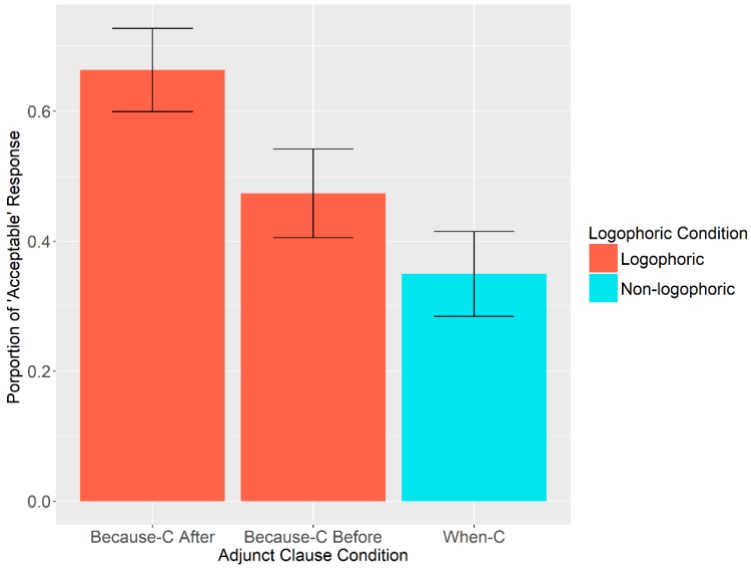


Fig. 6. Proportion of 'acceptable' responses in Adjunct Clause contrasted by logophoric condition (logophoric vs. non-logophoric). The error bars show 95% confidence interval.

3.4.4 Discussion

In sum, a significant improvement in acceptability of *taziji* was found in PP condition in which the antecedent is the perspective holder ('according to'). In addition, *taziji* was rated significantly more natural in adjunct clauses associated with mental activities of the antecedent ('because'-clauses) than in 'when'-clauses where there is less mental involvement.

Experiment 3 provided more supportive evidence that *taziji* can be exempt in various syntactic positions and can be LD bound across a local animate subject, contradicting some previous proposals. Though logophoricity can't capture all the obtained variance in acceptability, it appears to be more on the right track than the LDB theory for *taziji*. First, it's difficult for the LDB theory to explain why *taziji* can refer to the object of a Prepositional phrase which does not c-command it. Second, it is unclear how the LDB theory can capture the acceptability variance of *taziji* across different adjunct clauses ('because' vs 'when'), given that these structures are syntactically very similar.

Previous studies have argued that the referential properties of *ziji* and *taziji* should be accounted for with distinct mechanisms, Condition A for *taziji* and LDB/logophoricity theory for *ziji*. However, *ziji* and *taziji* are in fact much more similar than usually claimed: both of them are (i) exempt from Condition A, (ii) sensitive to logophoricity effects and (iii) morphologically complex. Hence, a more unified account of these two reflexives is potentially possible, which would simplify the current theories.

3.5 Summary of Experiments 2 & 3

Table 3: Implications to the theoretical claims of reflexive *taziji* in the literature

Theoretical issues	**Previous studies**	**Results of The Present study**
Referential property	Local anaphor (widely-accepted view)	Exempt from Condition A
C-commanding	Required (pure structural accounts)	Not necessary
Subject orientation	Unclear	No
Comparison with *ziji*	Binding condition of *ziji* and *taziji* are very different and should be derived by fundamentally different mechanisms.	*Ziji* and *taziji* are very similar, and their binding mechanisms might be categorically the same.

3.6 Conclusion

This paper focuses on disentangling the two major competing theories on exempt anaphors, LDB theory and logophoricity theory, via testing the referential properties of the two Chinese reflexives *ziji* and *taziji*. The results from Experiments 1-3 show that the distribution of z*iji* is better captured by the logophoricity theory than by the LDB theory. Furthermore, *taziji* is in fact more similar to *ziji* than usually claimed: *taziji* can take non-local antecedents and is also sensitive to logophoricity effects. Instead of proposing distinct theories on *ziji* and *taziji*, it might be simpler to model the two reflexives with consideration of discourse factors such as logophoricity.

This paper also casts doubt on the traditional view of correlating the morphology of reflexives and their binding conditions. Morphologically complex reflexives can be (non-strictly) local (English *himself*, Dutch *zichzelf*) or exempt (Chinese *ziji*, *taziji*, Japanese *zibun*, Dong *agen*). Morphologically simplex reflexives can also be (non-stricly) local (German *sich*) or exempt (Icelandic *sig*). It seems that there is no clear correlation between being mono/multi-morphemic and being exempt/local.

C-commanding-subject orientation of *ziji* (and potentially *taziji*) might be an artifact of logophoricity effects, since subjects are usually the perspective-takers of the whole clause. Indeed, both *ziji* or *taziji* were found to naturally refer to a non-c-commanding object NP in the present study.

In order to develop a more complete theory of the environments in which Chinese *ziji* and *taziji* are acceptable, there are still many open questions. One puzzle is whether local *ziji/taziji* are logophoric or not. Charnavel and Huang (2018) demonstrated that local inanimate *ziji* is not logophoric. More work needs to be done to investigate local animate *ziji*. Another related question is whether multiple *ziji* within the same clause can take different antecedents, which reflects how many different logophoric centers/domains a clause can have. Judgments for this issue in the literature are not uniform (cf: Pan 1997): Huang & Liu (2001) demonstrated that multiple occurrences of *ziji* in a clause can refer to distinct antecedents (a local one and a LD one), while Shuai, Gong & Wu (2013) report that in their experimental participants do not accept *ziji* referring to distinct antecedents within the same clause. Further work is clearly needed, perhaps including investigations of potential dialect differences.

Chapter 4: [Paper 3] Neighborhood candidate effects on noisy-channel processing

4.1 Introduction

Previous psycholinguistics studies have shown that comprehenders do not always interpret a sentence according to its literal meaning from the syntax. For instance, although *the cheese* should be the agent based on the form in (1a), around 15% of the responses from comprehenders regarded *the mouse* as the agent, when comprehenders were asked about the meaning of (1a) after hearing it (Ferreira, 2003). With the same paradigm, for (1b), around 10% responses treated *horse* as the agent, though it is actually the patient according to the syntactic rule (Ferreira, 2003). The compositional literal meaning in (1c) is that the candle is the recipient, and the daughter is being given to the candle. However, when comprehenders were ask about the meaning of (1c) after reading it, more than half of the responses corresponded to the plausible yet non-literal message – the daughter being the recipient and the candle being the transferred item (Gibson et al., 2013).

(1) a. The mouse was eaten by the cheese. (~ 15% non-literal interpretation)

b. The rock kicks the horse. (~ 10% non-literal interpretation)

c. The mother gave the candle the daughter. (~ 52% non-literal interpretation)

To capture the phenomena in (1a-c), three proposals have been proposed in the literature –the competition model (Bates et al., 1982), the good-enough parsing account (Ferreira, 2003), and the noisy-channel model (Gibson et al., 2013). Below We review these three previous accounts. We then introduce our proposed noisy-channel model based on the original model in Gibson et al. (2013) and the research questions it addresses.

4.1.1 The Good-enough Parsing Account

Ferreira (2003) proposed the 'good-enough parsing' account for (1a), copied below, which states that comprehenders sometimes adopt simple heuristics when parsing a sentence, in addition to (or instead of) accurate syntactic parsing algorithms that would lead to the literal (compositional) meaning of the sequence of words, resulting in a 'good enough' rather than accurate linguistic representation of an utterance (c.f. Christianson et al., 2006; Christianson, 2016; Slattery et al., 2013; Ferreira & Patson, 2007). The two heuristics involved in (1a) are the NVN strategy - a noun-verb-noun syntactic template is mapped into an agent-verb-patient thematic structure by the comprehender – and a semantic association, which prefers plausible meanings. This is proposed as an explanation for why comprehenders sometimes reach an erroneous representation of the input sentence and regard the literal patient *mouse* as the agent (Ferreira, 2003).

(1) a. The mouse was eaten by the cheese. (~ 15% non-literal interpretation)

A limitation of this proposal is that it does not provide a detailed account for how comprehenders reach the non-literal interpretation across different kinds of constructions. For example, the heuristics for 'misinterpreting' the sentence formed in dative structure (1c) *The mother gave the candle the daughter* might be different from (1a), as there are three Ns with three different kinds of thematic roles - and how the heuristics are combined with each other to reach the final (mis)interpretation. These issues make it difficult for this account to generalize to a wider range of constructions.

4.1.2 The Competition Model

A second account is provided by the competition model (e.g., Bates et al., 1982; MacWhinney, 1977, 1992, 2005, 2010; Li & MacWhinney, 2013), which proposes that language learners/users link linguistic form with meaning via various *cues.* A cue is essentially an information source coming from morphology, syntax, or semantics, or arbitrary combinations of each. An utterance always instantiates many cues, and these cues may converge or compete, with the result that cues can point in the same or different directions for meaning interpretation in a sentence. In case of competition, the weights that comprehenders assign to different cues determine the final interpretation of the sentence (Li & MacWhinney, 2013). This account was proposed as a way to capture language processing and language learning in a general framework. For instance, in simple sentences that consist of two NPs and a verb, such as (1b) where the English word order cue indicates *the rock* is the agent, while the semantics cue suggests *the horse* is the agent, thus the competition between these two cues leading to distinct interpretations.

(1) b. The rock kicks the horse. (~ 10% non-literal interpretation)

4.1.3 The Noisy-channel Model

A recent more general account of 'misinterpretation' effects is provided by the noisy-channel framework (e.g., Levy, 2008; Gibson et al., 2013). Given that noise is present in typical language use, such as environment noise, speech errors, printing errors and many others, the noisy-channel model was proposed to capture how comprehenders successfully figure out what the speaker wants to convey, when the utterance is likely corrupted by noise. From the perspective of the noisy-channel model, the non-literal interpretations ('misinterpretations') come from comprehenders' rational inferences about the speaker's intended utterance which might be different from the perceived input sentence in the context of noise. As in (2), recent results suggest that comprehenders do so by rationally combining the prior

probability of the possible intended utterance based on prior knowledge $P_L(s_i)$, with a noise model $P_N(s_i \rightarrow s_p)$ which encodes how an intended utterance S_i might be corrupted to a perceived sentence S_p in transmission.

(2) $P(s_i|s_p) \propto P_L(s_i)P_N(s_i \rightarrow s_p)$

The prior, $P_L(s_i)$ represents the comprehender's relevant linguistic and world knowledge, and biases comprehenders towards a priori plausible utterances. The noise model $P_N(s_i \rightarrow s_p)$ encodes the comprehender's knowledge of how sentences can be corrupted - for instance, smaller changes to a sentence are more likely than larger ones. By integrating $P_L(s_i)$ and $P_N(s_i \rightarrow s_p)$, comprehenders may arrive at interpretations which differ from the literal meanings of the input sentence. In another word, comprehenders engage in optimal Bayesian decoding of the intended meaning via maximizing $P(s_i|s_p)$ – the probability of the intended utterance given the perceived sentence (e.g., Levy, 2008; Bergen et al., 2012; Gibson et al., 2013; Poppels & Levy, 2016; Ryskin et al., 2018).

For instance, if comprehenders perceive an input sentence (1c), copied below, comprehenders might infer that an error occurred between the speaker's intent and the ultimately perceived sentences, such that the intended sentence was actually the more plausible (1d). From the comprehender's perspective, the perceived sentence (1c) has an implausible meaning with a very low semantic prior. Given that its neighborhood candidate (1d) has a much higher semantic prior than the input (1c), and these two sentences are reasonably similar with just one word deletion from (1d) to (1c). As deletion is a relatively common error which only requires a word to be selected from a sentence, it is very likely that the comprehenders infer that a deletion error occurred in transmission, and the speaker's intended message

is the more plausible (1d) (Gibson et al., 2013). That explains why there were over 50% non-literal interpretations for (1c) in previous experiments.

(1) c. *Perceived:* The mother gave the candle the daughter. (~ 48% literal interpretation)

d. *nearby sentence :* The mother gave the candle to the daughter. (deletion error)

One major difference between the good enough parsing account and the noisy-channel model is that the good enough parsing account does not take into account the neighborhood candidates - the sentences that are very close to the perceived sentence, such that they might be the actual intended utterance. This lack of consideration of neighborhood candidate effects makes it hard for the good enough parsing account to generalize to a wider variety of constructions, because it does not set a boundary for the involved heuristics for each new construction. In addition, neither the good enough parsing nor the competition model considers corruption due to all sorts of noise in the information transmission procedure.

Previous research in the noisy-channel framework focused on the noise model and the world-knowledge prior (Gibson et al., 2013). An important limitation of this early work was that it did not explore the potential existence of a language prior (cf. Bergen et al., 2012; Poppels & Levy, 2016; Keshev & Meltzer-Asscher, 2021). A second limitation with earlier noisy-channel work was that it focused mostly on English examples (but cf. Keshev & Meltzer-Asscher, 2021; Zhan et al, 2018; Poliak et al., 2022). In this work, we seek to explore the syntactic prior in English and in Mandarin.

4.1.4 Our Proposed Noisy-channel Model with Integration of Syntactic Information

Previous works of the noisy-channel model were built upon the assumption that neighborhood candidates play a role in processing possibly corrupted sentences. However, in-depth investigation of the structural properties of the input and its neighborhood candidates is missing.

There have been several pieces of evidence in the literature suggesting the rate of non-literal interpretation is affected by the frequency of the form (cf. Bergen et al., 2012; Poppels & Levy, 2016; Keshev & Meltzer-Asscher, 2021). For instance, Poppels and Levy (2016) showed that participants were more likely to interpret a sentence non-literally from a low frequency preposition phrase ('to…from'), such as 'The package fell **to** the table **from** the floor', than a sentence from a high-frequency preposition phrase ('from…to'), such as 'The package fell **from** the floor **to** the table.'

Pace the previous works, we define the language prior $P_L(s_i)$ as the joint probability of the meaning prior $P_L(s_{i\_meaning})$ and the structure prior $P_L(s_{i\_grammar})$, as in (3) and (4). According to our proposed noisy-channel model (4), both meaning plausibility and structural frequency affect comprehenders' interpretation of the input sentence. Sentences with low structural prior can trigger more non-literal interpretations, as the neighborhood candidates with higher structural prior might be considered as the actual intended utterance.

(3) $P_L(s_i) = P_L(s_{i\_grammar}, s_{i\_meaning})$

(4) $P(s_i|s_p) \propto P_L(s_{i\_grammar}, s_{i\_meaning})\, P_N(s_i \rightarrow s_p)$

In the following sections, we will walk through the three major research questions this paper addresses with respect to our proposed noisy-channel model with implementation of the structural prior.

A. How the structural prior works

Our proposed model (4) predicts that comprehenders are more likely to draw inferences for sentences formed in low frequency structure than those formed in high frequency structures. The reason is that comprehenders are more likely to consider that an error occurred in the input sentence with infrequent structures, inferring the neighborhood candidates with a higher structural prior as the actual intended message from the speaker.

The specific construction under consideration here is simple transitives: For input sentences of the same implausible meaning (in terms of predicate-argument relationship) formed in frequent SVO (5a) or infrequent OSV (6a) structures, (6a) has at least three neighborhood candidates of higher structural or/and meaning prior (6b-d) that could attract the comprehender, while only (5b) has a higher language prior than (5a) among (5b-d). That's why (6a) is predicted to result in more non-literal interpretations than (5a) by our noisy-channel model.

(5)	SVO	The language prior $P_L(s_{grammar}, s_{meaning})$	Compared to $P_L(s_p)$
	a. *Perceived S_p:* The apple ate the boy.	$P_L(s_{p\_SVO}, s_{p\_implau})$	
	b. *Nearby S_1*: <u>The boy ate the apple.</u>	$P_L(s_{1\_SVO}, s_{1\_plau})$	**Higher**
	c. *Nearby S_2*: The apple, the boy ate.	$P_L(s_{2\_OSV}, s_{2\_plau})$	Might be lower
	d. *Nearby S_3*: The boy, the apple ate.	$P_L(s_{3\_OSV}, s_{3\_implau})$	Lower

(6)	OSV	The language prior $P_L(s_{grammar}, s_{meaning})$	Compared to $P_L(s_p)$
	a. *Perceived S_p:* The boy, the apple ate.	$P_L(s_{p\_OSV}, s_{p\_implau})$	
	b. *Nearby S_1*: <u>The boy ate the apple.</u>	$P_L(s_{1\_svo}, s_{1\_plau})$	**Higher**
	c. *Nearby S_2*: <u>The apple, the boy ate.</u>	$P_L(s_{2\_osv}, s_{2\_plau})$	**Higher**
	d. *Nearby S_3*: <u>The apple ate the boy.</u>	$P_L(s_{3\_svo}, s_{3\_implau})$	**Higher**

B. The 'grain sizes' distributional syntactic information in the structural prior

A key question about the structural prior in our noisy-channel model is what kind of frequencies comprehenders are tracking. Previous works have demonstrated that comprehenders are sensitive to prior linguistic experience in the sense of probabilistic context-free grammar rules, where the probabilities of all the constituents in a sentence are counted (e.g., surprisal in Hale (2001) and Levy (2008)), but it is not yet known what 'grain sizes' distributional syntactic information is stored by language users (e.g., Mitchell et al., 1995). In another word, we need to figure out the 'boundary' for the rational inference.

Here we evaluate two hypotheses about the level of syntactic information tracked by comprehenders, as in Table 1 below. One is the construction-based hypothesis: comprehenders track constructions – for instance, comprehenders treat simple transitives and clefts as separate constructions which contain different structures (SVO/OSV or subject/object clefts) (Goldberg, 2016; Abeillé et al., 2020). Another is the linear string-based hypothesis: the comprehension mechanism tracks the linear string of content

words – for example, NVN for SVO and subject cleft, and NNV for OSV and object cleft (Bates et al., 1982; Ferreira 2003).

Table 1. Illustration of the construction-based hypothesis and the linear string-based hypothesis.

Pair	**Sentence**	**Construction-based hypothesis**	**Linear-string based hypothesis**
Simple transitives	a. The trash threw the boy.	Simple transitive (SVO)	NVN
	b. The boy, the trash threw.	Simple transitive (OSV)	NNV
Clefts	c. It was the trash that threw the boy.	Cleft (Subj)	NVN
	d. It was the boy that the trash threw.	Cleft (Obj)	NNV

These two hypotheses make distinct predictions for comprehenders' interpretation of (implausible) sentences formed in simple transitives and clefts. The construction-based hypothesis predicts that there should be variation between simple transitives and clefts in the amount of non-literal responses from comprehenders, since simple transitives and clefts are tracked as separate constructions (Goldberg, 2016; Abeillé et al., 2020). On the other hand, the linear string-based hypothesis predicts no variation between simple transitives and clefts, as they contain the same pair of strings – NVN and NNV (Bates et al., 1982; Ferreira 2003). For more details about the predictions of the two hypotheses, please see Section 4.4.1.

C. Where the non-literal interpretations come from

Though the discussion of non-literal interpretations can be dated back to at least Bates et al. (1982), where these non-literal interpretations come from still remains unclear.

Here we define two competing hypotheses about the source of the non-literal responses. One is our hypothetical speaker's channel hypothesis. According to this hypothesis, comprehenders are fully aware of the input sentence, and the non-literal responses result from their rational inferences about the speaker's intended utterance which might be corrupted due to the speaker or in the transmission procedure. The alternative is comprehender's channel hypothesis - comprehenders are *not* fully aware of the input sentence, due to their mis-reading/mis-retrieval of the input, so that they reach an 'incorrect' representation (e.g., Ferreira, 2003). To distinguish these two accounts, we added a retyping task in Experiment 5 to probe whether comprehenders truly mis-represent the input sentence.

The remainder of this paper is structured as follows. Experiments 1-2 quantitatively measured the degree to which English and Mandarin Chinese allow the six logically possible word orders of a subject (S), verb (V) and object (O): SVO, OSV, SOV, VOS, OVS, VSO. A corpus search was also conducted to estimate the frequencies of the structures in natural production. Experiments 3-4 tested our proposed noisy-channel model with consideration of the structural prior with simple transitives in English and Mandarin. Experiment 5 further tested our proposed noisy-channel model beyond simple transitives – both (subject/object) clefts and simple transitives were tested. In addition, Experiment 5 was also designed to evaluate the two hypotheses about distributional syntactic information stored by language users (construction-based vs. linear string-based) and the two competing accounts of where the non-literal interpretations come from (hypothetical speaker's channel vs. comprehender's channel). To make sure our findings are robust and replicable, Experiment 6 was conducted as a replication of Experiment 5.

4.2 Experiments 1-2

Experiments 1-2 aimed to experimentally measure and compare the degree to which English and Mandarin Chinese allow the six logically possible word orders (SVO, OSV, SOV, VOS, OVS, VSO) for simple transitive events. The two experiments examined the acceptability of the six word orders, as an approximate measure of the frequencies of these word orders in the two languages.

To our knowledge, Experiments 1-2 offer the first experimental design that allows us to quantify word order flexibility in various languages and compare them cross-linguistically.

4.2.1 Participants

30 English speakers and 30 Mandarin speakers participated in Experiment 1 and 2 respectively. English speakers were recruited via Amazon Mechanical and Mandarin speakers via Witmart.

4.2.2 Design and Materials

The experiment included 8 test items and 12 fillers. In the test trials, participants were presented with pictures depicting various events and the six logically possible word orders for each picture, as in (7). A comma was inserted between two consecutive NPs to avoid ambiguity, which would be especially problematic for Mandarin. Participants were asked to select all the acceptable descriptions among the six word orders.

The 12 fillers had from 1 to 6 clearly correct answers among the 6 descriptions. An example filter item with 4 correct answers is provided in (8).

The test and filler items were presented with a random order for each of the 30 participants. In each trial, the order of the six sentences was randomized. In this way, participants would be encouraged to read all the 6 sentences and choose all the appropriate depictions among them.

(7) Sample test item

Please select ALL the acceptable English description(s) of the picture.

The boy threw the trash.	(SVO)
The trash, the boy threw.	(OSV)
Threw the trash, the boy.	(VOS)
The boy, the trash threw.	(SOV)
The trash threw the boy.	(OVS)
Threw the boy, the trash.	(VSO)

(8) Sample filler item with 4 correct answers

Please select ALL the acceptable English description(s) of the picture.

Beside the boy, there is a car.	(correct)
Beside the car, there is a boy.	(correct)
There is a car beside the boy.	(correct)
There is a boy beside the car.	(correct)
Beside the boy, there is a train.	(wrong)
Beside the ship, there is a boy.	(wrong)

4.2.3 Results

Data from participants who did not self-report as native speakers of American English/Mandarin Chinese or didn't answer all the filler trials with at least 75% accuracy were excluded.

Responses from 29 English-speaking and 30 Chinese-speaking participants were analyzed. Two mixed-effects logistic regressions with word order as predictor and subject as the random intercept[23] were applied to the English and the Chinese datasets separately. The results showed that SVO was significantly more frequently chosen than OSV, the second preferred word order, in English (β=8.1, z=8.2, p<0.001) as well as in Mandarin (β>20, z>10, p<0.001), while other word orders are less common (Figs. 1 & 2).

Furthermore, the entropy in the distribution of allowable Mandarin word orders (H=1.28, 95% CI=[1.26, 1.3]) was larger than that in English (H=0.95, 95% CI=[0.9, 1]), suggesting Mandarin has more flexible word orders than English.

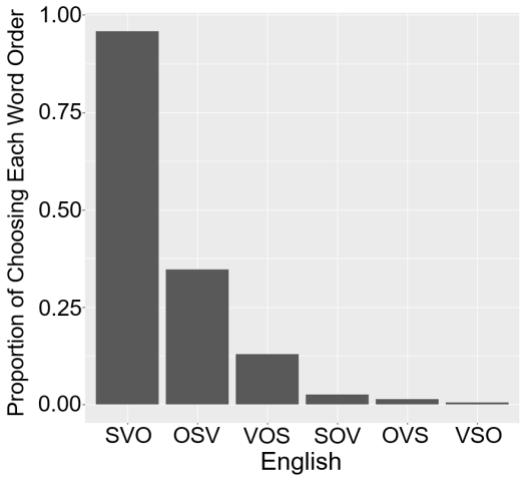


Fig.1. Proportion of choosing the six logically possible word orders in English.

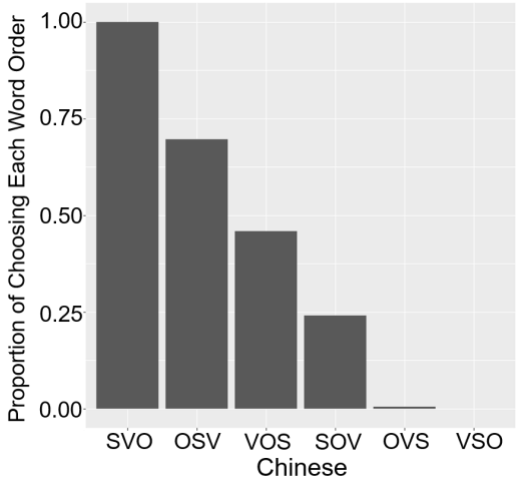


Fig.2. Proportion of choosing the six logically possible word orders in Mandarin.

4.2.4 Corpus search

[23] This is the maximal model that can converge for both datasets.

A corpus search in the English and Mandarin Penn Treebanks showed that while SVO was more common than OSV (topicalization) in both languages, the proportion of OSV was higher in Mandarin, as in Table 2 below. Such results align with our findings from Experiments 1 and 2.

Table 2[24]:

	Raw counts		Probability of each structure	
	SVO	OSV	SVO	OSV
English	87515	176	0.62	0.001
Chinese	65688	1693	0.64	0.015

4.2.5 Discussion

The results of Experiments 1 & 2 demonstrate that (a) SVO is a more frequent/acceptable word order than OSV in both languages; and (b) Mandarin has more flexible word orders than English based on experimental data.

The relative acceptability of OSV to SVO in Mandarin (Fig.3) seems to be much higher than its relative frequency in corpus search (Table.1), which might be (partially) due to the difference between casual spoken Mandarin - as in the experiment - and formal written text (mainly news) in the searched corpora. Another possible reason might be that the experiment probes people's acceptability of sentences, which may not be linearly correlated with their frequency of using these word orders in natural production (as reflected by the corpora).

[24] Both the raw counts and the probability metrics were provided by the Penn Treebanks. The denominator of the probability of each structure should be all the structures (not just simple transitives) in English and Chinese identified by the Penn Treebanks.

4.3 Experiments 3-4

The goal of Experiments 3 and 4 was to test our proposed noisy-channel model with consideration of the structural prior. Experiments 3 and 4 measured English and Chinese speakers' comprehension of implausible and plausible sentences formed in the most frequent SVO word order or the second most frequent OSV.

As stated in the Introduction, our proposed noisy-channel model predicts more non-literal responses for low-frequency OSV than high-frequency SVO, as OSV has a lower structural prior and comprehenders are more likely to infer that the intended utterance is not the input sentence but its neighborhood sentences.

It is vital to test the proposed noisy-channel model with non-English languages, in this case, Mandarin Chinese, especially given that there hasn't been much cross-linguistic study of noisy-channel processing.

4.3.1 Participants

107 English speakers and 87 Mandarin speakers participated in Experiment 3 and 4, respectively. English speakers were recruited via Amazon Mechanical and Mandarin speakers via Witmart.

4.3.2 Design and Materials

12 test items were constructed following a 2x2 design – plausibility (implausible/plausible meaning) and word order (SVO/OSV). Each sentence was paired with a comprehension question that distinguished literal and non-literal interpretations (9). The 12 items were split across 4 lists: each list contained 3 sentences for each of the 4 conditions. There were also 24 filler items including comprehension questions with clearly correct answers in each list.

(9) The trash threw the boy. (SVO_implausible)
The boy, the trash threw. (OSV_implausible)
The boy threw the trash. (SVO_plausible)
The trash, the boy threw. (OSV_plausible)

Question: Did the boy throw something/someone?

4.3.3 Results

Data from subjects who were not native speakers of American English or Mandarin Chinese were excluded. We also filtered out those who did not answer all the comprehension questions of the fillers with at least 85% accuracy. Responses from 97 English speakers and 81 Mandarin speakers were analyzed**.**

Two mixed-effects logistic regressions were applied to the English and the Mandarin datasets separately. For data from English speakers, we entered word order and plausibility as the predictors, and by-subject and by-item random intercepts.[25] The results supported the structural prior hypothesis - OSV sentences were significantly more likely to be interpreted non-literally than SVO sentences in English (β = 2.36, z=8.22, p<0.001). There was a significantly higher amount of non-literal responses (β = 1.95, z=7.39, p<0.001) for implausible sentences than for plausible sentences (see Figure 3).

[25] In the rest of the paper, we first fit the model with the maximal random effect structure. We reported results of this model, if it converged. Otherwise, we started from the random intercepts-only model, including random intercepts for subject and item, and performed forward model comparison using likelihood-ratio tests (anova function in R). Random slopes will be included in the final model, if model fitness is significantly improved. In cases where we needed to choose between two non-nested random effect structures and likelihood-ratio test cannot be deployed, we reported the model with a lower Bayesian Information Criterion (BIC).

For the Mandarin data, word order and plausibility were entered as the predictors. The model was fit with random by-subject and by-item intercepts as well as random slope for plausibility by-subject. As predicted by the structural prior proposal, Mandarin speakers were significantly more likely to interpret OSV sentences non-literally than SVO sentences (β = 1.88, z=6.76, p<0.001). In addition, implausible sentences were more likely to be interpreted non-literally than plausible sentences (β = 2.7, z=3.16, p<0.01) (see Figure 4).

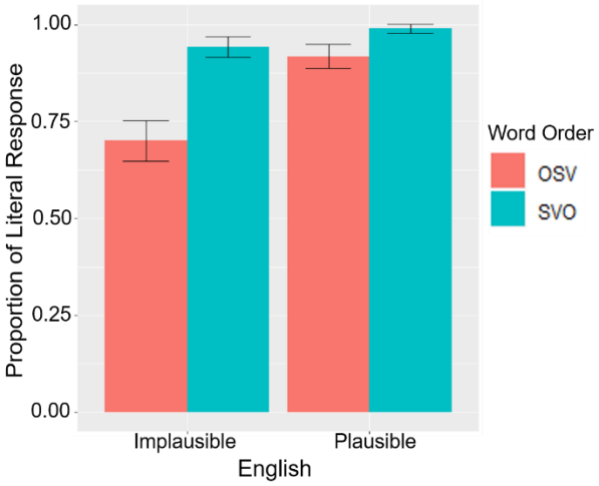


Figure.3. Proportion of choosing literal interpretation by word order and plausibility in English (95% CI).

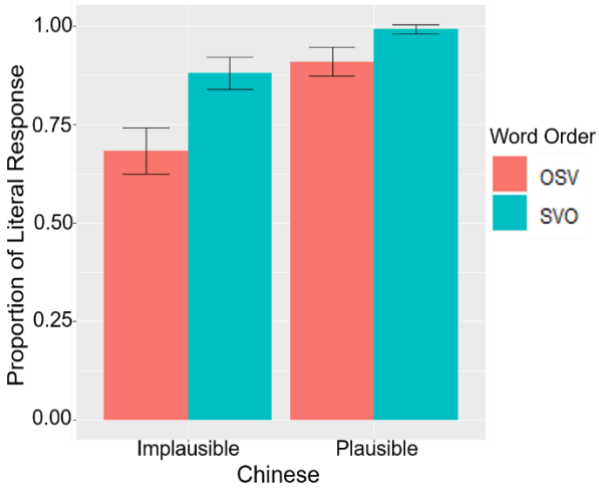


Figure.4. Proportion of choosing literal interpretation by word order and plausibility in Mandarin (95% CI).

4.3.4 Discussion

The results showed that OSV sentences were more likely to be interpreted non-literally compared to SVO sentences in both English and Mandarin. Thus, as predicted by our proposed noisy-channel model, both English and Mandarin speakers are much more likely to interpret "The boy, the trash threw" as the more plausible "The boy threw the trash" than they are to interpret "The trash threw the boy" in the more plausible way.

Results from Experiments 1 – 4 have demonstrated that comprehenders are sensitive to the structural prior of the input sentence. But there are two important remaining questions: The first puzzle is at what 'grain sizes' distributional syntactic information is tracked by language users (e.g., Mitchell et al., 1995). And the second question is where the non-literal responses come from. To answer these two questions, Experiment 5 was conducted to test both simple transitives (SVO and OSV) and clefts (subject and object clefts) with a novel experiment design.

4.4 Experiment 5

4.4.1 Overview of the experiment

A. Goals

Experiment 5 had three goals. First, it aimed to replicate our findings in Experiment 3 about simple transitives.

Second, Experiment 5 was designed to further test our proposed noisy-channel model beyond simple transitives and evaluate the two hypotheses about the distributional syntactic information stored by language users: the construction-based vs. the linear string-based hypothesis. We therefore tested subject and object clefts, in addition to simple transitives.

The third goal of this experiment was to investigate where the non-literal interpretations come from – more specifically, to tease apart the hypothetical speaker's channel hypothesis and the comprehender's channel hypothesis.

B. Theories and predictions

1. Our proposed noisy-channel model with structural prior

Our proposed noisy-channel model with integration of the structural prior predicts that low frequency structures should trigger more non-literal interpretations than high-frequency structures, if these structures are neighborhood candidates for each other. Given that object clefts are less frequent than subject clefts (Roland et al., 2009), our noisy-channel model predicts more inferences for object clefts than subject clefts. Thus, we expect to (i) replicate the results of Experiment 3, finding more non-literal responses for OSV than SVO, and (ii) observe more non-literal interpretations for object clefts than subject clefts, as in Figure 5.

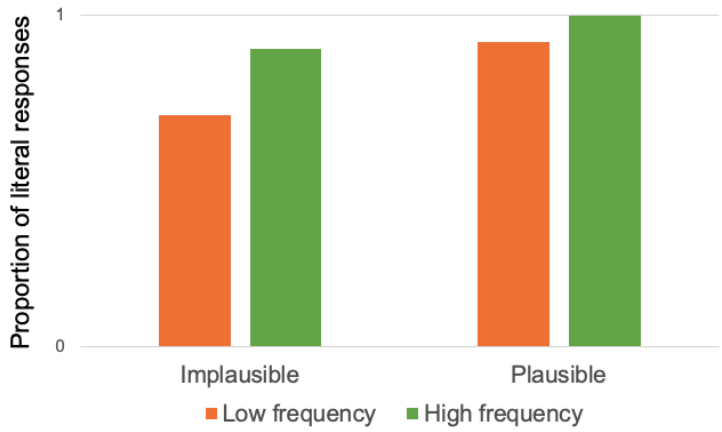


Figure 5. Prediction of our proposed noisy-channel model for low- and high- frequency structures.

2. Distributional syntactic information tracked by comprehenders

Delving deeper into the implemented structural prior in the noisy-channel model, a key question is what kind of frequencies comprehenders are tracking – in another word, what 'grain sizes' distributional syntactic information is stored by language users (e.g., Mitchell et al., 1995).

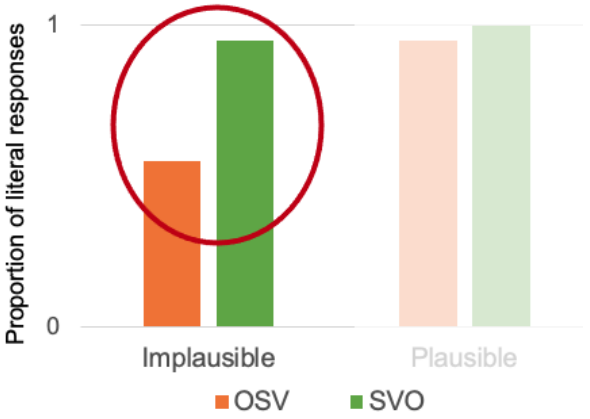
We proposed two hypotheses about the level of syntactic information tracked by comprehenders, based on previous literature, as in Table 3 below. One is the construction-based hypothesis: comprehenders track constructions – they treat simple transitives and clefts as separate constructions which contain different structures (SVO/OSV or subj/object clefts) (Goldberg, 2016; Abeillé et al., 2020). Another is the linear string-based hypothesis: the comprehension mechanism tracks the linear string of content words – NVN for SVO & subject cleft, and NNV for OSV and object cleft (Bates et al., 1982; Ferreira 2003).

Table 3. Illustration of the construction-based hypothesis and the linear string-based hypothesis.

Pair	**Sentence**	**Construction-based hypothesis**	**Linear-string based hypothesis**
Simple transitives	a. The trash threw the boy.	Simple transitive (SVO)	NVN
	b. The boy, the trash threw.	Simple transitive (OSV)	NNV
Clefts	c. It was the trash that threw the boy.	Cleft (Subj)	NVN
	d. It was the boy that the trash threw.	Cleft (Obj)	NNV

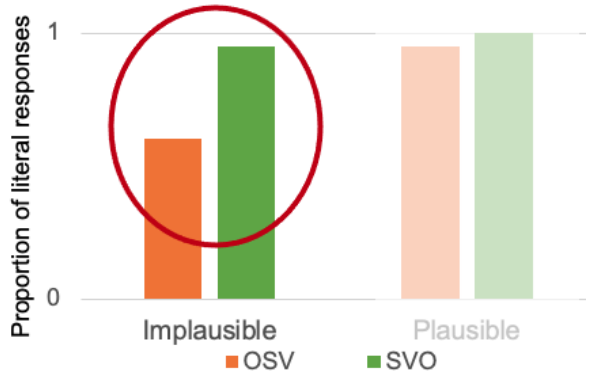
These two hypotheses make distinct predictions for comprehenders' responses. The construction-based hypothesis predicts that (i) the overall inference rate should differ between simple transitives and clefts, as they are distinct constructions; and (ii) the difference in inference rate is larger between SVO and OSV than between the two types of clefts, as in Fig. 6ab. The relative frequency contrast between SVO vs. OSV is much larger than that between subject vs. object clefts - the raw counts ratio between SVO and OSV is 497: 1, while the ratio between subject vs. object cleft is 17:1, based on our corpus search in the English Penn Treebank and Roland et al. (2007). The amount of inference should be in proportion to the relative frequency contrast within each construction, as the more frequent the input, the less likely the comprehenders will be biased towards the neighborhood candidates.

On the other hand, the linear string-based hypothesis predicts that the inference rates for simple transitives and clefts should be similar, and the inference rate difference between SVO and OSV should be similar to that between subject and object clefts, as the processing mechanism treats OSV and object cleft the same as NNV, SVO and subject cleft as NVN, as in Fig. 6cd.

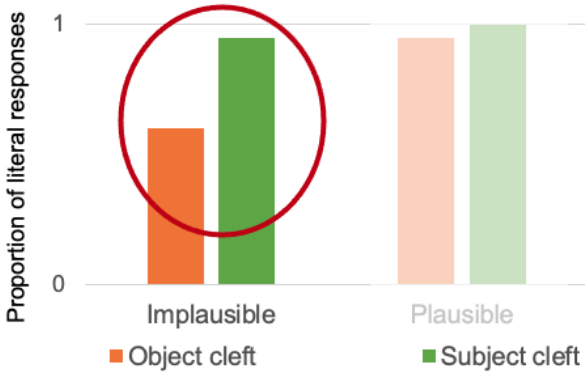


a. Prediction of the construction-based hypothesis for simple transitives.

b. Prediction of the construction-based hypothesis for clefts.



c. Prediction of the linear string-based hypothesis for simple transitives.



d. Prediction of the linear string-based hypothesis for clefts.

Figure 6. Predictions of the comprehension question results based on the construction-based hypothesis (a-b) and the linear string-based hypothesis (c-d). (Responses of the plausible conditions are almost always near the ceiling and therefore de-emphasized in this figure)

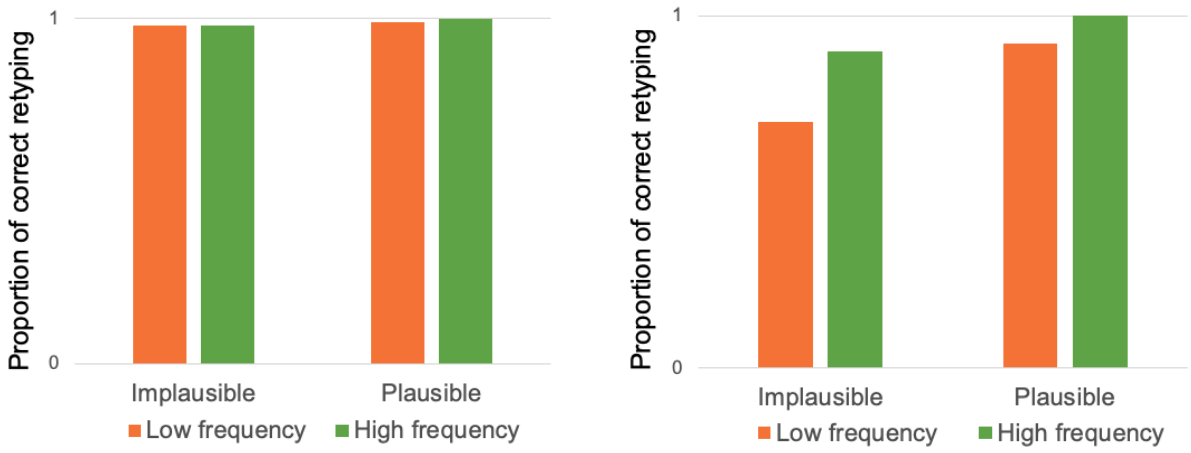
The construction-based and the linear string-based hypotheses also set distinct 'boundaries' for the inference – to what extent a structure can be corrupted into another one. The construction-based hypothesis proposes that distinct constructions (simple transitives vs. clefts) should treated differently by the comprehender, and that corruption happens only among neighborhood candidates that are sufficiently similar to the input sentence. The linear string-based hypothesis, on the other hand, states that different constructions can be regarded as the same by comprehenders, as long as they share the same pair of strings, such as OSV and object cleft.

3. Where the non-literal interpretations come from

Though the discussion of non-literal sentence comprehension can be dated back to at least Bates et al. (1982), where the non-literal interpretations come from still remains unclear. Indeed, there has been no falsifiable theories that make clear predictions for experimental data. We defined two falsifiable competing hypotheses about the source of the non-literal responses which can be tested by our novel experimental design. One is our hypothetical speaker's channel hypothesis - comprehenders are fully aware of the input sentence. The non-literal responses result from their rational inferences about the speaker's intended utterance which might be corrupted due to the speaker or in the transmission procedure. The alternative is comprehender's perceptual channel hypothesis - comprehenders are *not* fully aware of the input sentence, due to their misreading/mis-retrieval of the input, so that they reach an 'incorrect' interpretation (c.f., Ferreira, 2003).We added a retyping task in Experiment 5 to evaluate these two accounts.

Our proposed hypothetical speaker's channel hypothesis predicts almost no incorrect retyping, because comprehenders are fully aware of the input sentence, as in Figure 7a. In contrast, the comprehender's

channel hypothesis predicts the proportion of incorrect retyping align the amount of non-literal responses, as the non-literal interpretations result from misperception, in Figure 7b.



a. Prediction of the hypothetical speaker's channel hypothesis

b. Prediction of the comprehender's channel hypothesis

Figure 7. Predictions of the hypothetical speaker's channel and the comprehender's channel hypothesis.

C. Preview of the experiment

Experiment 5 included two sub-experiments: (i) Experiment 5a is a replication and extension of Experiment 3, testing English simple transitives (SVO vs. OSV) with a different paradigm – a retyping task was added right after the comprehension question in each trial. (2) Experiment 5b, which had the same design as Experiment 5a, further tested our proposal on another construction – subject and object clefts.

4.4.2 Participants:

60 and 78 native English speakers were recruited via Prolific in Experiment 5a and 5b respectively.

4.4.3 Design and Materials:

12 test items were constructed following a 2x2 design – plausibility (implausible vs. plausible meaning) and construction type (SVO vs. OSV or subject vs. object clefts as in (10)). Each sentence was paired with a comprehension question that distinguished literal and non-literal interpretations, as in Experiments 3 & 4. The 12 items were split across 4 lists: each list contained 3 sentences for each of the 4 conditions. There were also 38 filler items in each list.

(10) It was the trash that threw the boy. (subjcleft_implausible)
It was the boy that the trash threw. (objcleft_implausible)
It was the boy that threw the trash. (subjcleft _plausible)
It was the trash that the boy threw. (objcleft _plausible)

Question: Did the boy throw something/someone?

In both experiments, participants read the input sentence either word by word (clefts) or phrase by phrase (simple transitives) with a self-paced reading paradigm and do two tasks for each sentence: (i) answer a comprehension question about the meaning of the sentence (same as in Experiments 3-4); (ii) retype the input sentence exactly in the form they read.

4.4.4 Results

Data from subjects who were not native speakers of American English or who did not answer all the comprehension questions in the fillers with at least 85% accuracy were excluded. Responses from 50 participants in Experiment 5a and 72 participants in Experiment 5b were analyzed.

A. The comprehension question task

Responses were analyzed with a mixed-effects logistic regression using the *lme4* package in *R.*

For simple transitives tested in Experiment 5a, the model was fit with word order (SVO vs. OSV) and plausibility (plausible vs. implausible) as predictors, as well as by-subject and by-item random intercepts. The results replicated Experiment 3 – low-frequency OSV sentences were more likely to be interpreted non-literally compared to high-frequency SVO sentences ($\beta=3.25$, $z=6.47$, $p<0.001$). There were also significantly more non-literal responses for implausible sentences than plausible sentences ($\beta=4.13$, $z=6.76$, $p<0.001$) as in Fig. 8.

For the two types of clefts tested in Experiment 5b, as predicted by our proposed noisy-channel model, low-frequency object clefts were significantly more likely to be interpreted non-literally compared to high-frequency subject clefts ($\beta=1.44$, $z=3.39$, $p<0.001$). In addition, similar to our findings with simple transitives, implausible sentences were more likely to be. Interpreted non-literally ($\beta=4.28$, $z=3.93$, $p<0.001$) , as in Fig. 9.

To tease apart the construction-based hypothesis and the linear string-based hypothesis, we analyzed responses of the implausible condition for both simple transitives and clefts. We entered frequency (high vs. low), construction (simple transitives vs. clefts), and their interaction as the predictors, as well as by-subject and by-item random intercepts. As predicted by the construction-based hypothesis, we found a significant interaction between frequency and construction - the difference in inference rates between SVO vs. OSV is larger than that between subject and object clefts ($\beta=-2.05$, $z=-2.8$, $p<0.01$). Indeed, the amount of inference made by comprehenders was in proportion to the relative frequency contrast between the input sentence and its neighborhood candidates. We also observed a main effect of construction ($\beta=-1.24$, $z=-2.08$, $p<0.04$), suggesting overall there were more non-literal responses for simple transitives than clefts. The results supported the construction-based hypothesis, not the linear string-based hypothesis. In addition, consistent with our previous findings, low frequency structures

(OSV & object clefts) were more likely to be interpreted non-literally than high-frequency ones (SVO & subject clefts) (β=-2.61, z=-6.65, p<0.001).

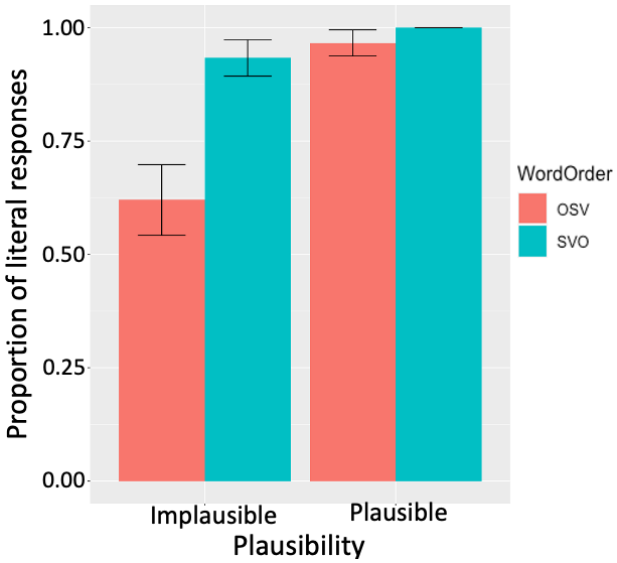


Fig.8. Proportion of choosing literal interpretation by word order (SVO vs. OSV) and plausibility in English (95% CI).

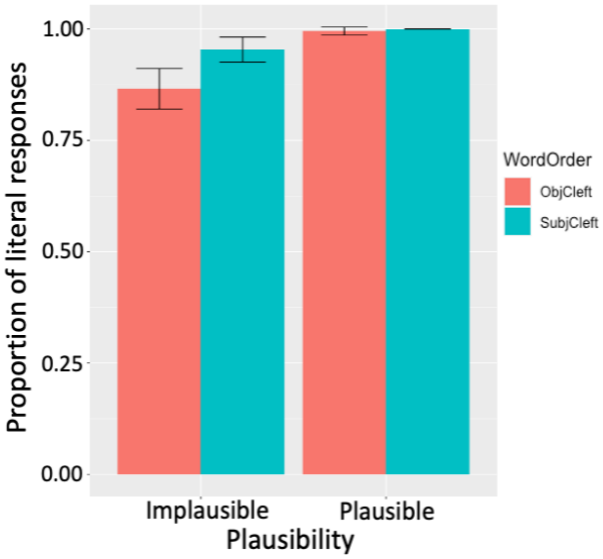


Fig.9. Proportion of choosing literal interpretation by structure (subject vs. object clefts) and plausibility in English (95% CI).

B. The retyping task

In the retyping task, the proportion of non-literal responses do not align the amount of incorrect retyping for both simple transitives and clefts. Almost all the participants (>99% of total responses) accurately retyped the exact form of the input sentence, including those who provided non-literal responses to the comprehension questions, as in Figs. 10 & 11. These results are consistent with our hypothetical speaker's channel hypothesis, contradicting the comprehender's channel hypothesis (c.f., Ferreira, 2003).

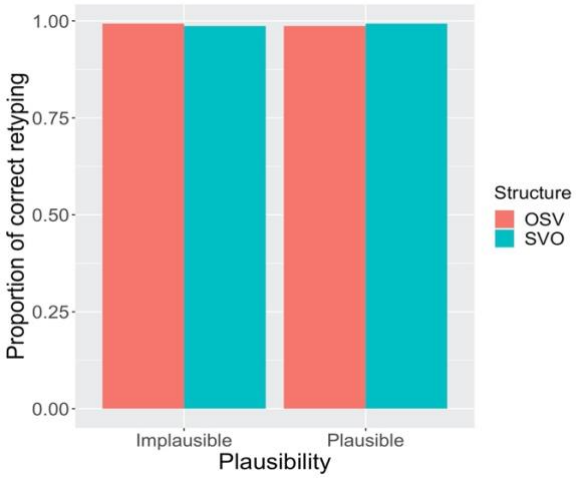


Fig.10. Proportion of correct retyping by word order (SVO vs. OSV) and plausibility in English (95% CI).

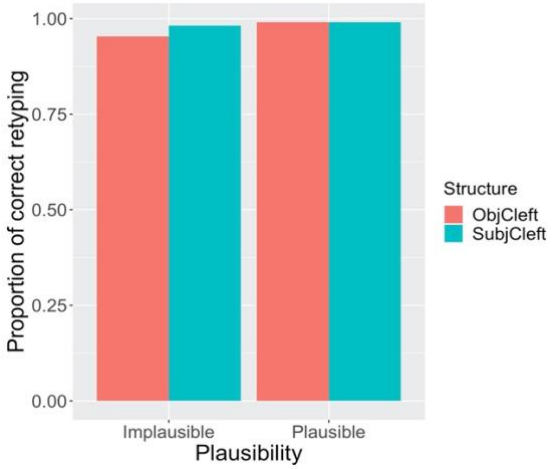


Fig.11. Proportion of correct retyping by structure (subject vs. object clefts) and plausibility in English (95% CI).

4.4.5 Discussion:

In Experiment 5, the observation of more non-literal interpretations for low-frequency structures than for high-frequency structures was replicated in simple transitives and extended to clefts, which provides robust supportive evidence for our structural frequency hypothesis.

The significant interaction between frequency (low vs. high) and construction (simple transitives vs. clefts) shows the amount of inference is in proportion to the relative frequency contrast between the input and its neighborhood candidates, supporting the construction-based hypothesis, not the linear string-based hypothesis. Comprehenders track each construction separately, not just the linear string. An input sentence can only be recovered to its neighborhood candidates which are sufficiently similar to it. For instance, an object cleft might be recovered to a subject cleft, but not to a dative structure or an OSV structure.

The finding that there was almost no incorrect retyping for almost all the responses, including the non-literal interpretations, suggests comprehenders' non-literal responses do not come from misperception of

the input sentence. Thus, there is no evidence for the comprehender's channel hypothesis. These results are consistent with our proposed hypothetical speaker's channel hypothesis.

All the three findings above align the predictions of our proposed noisy-channel model with integration of syntactic information.

4.5 Experiment 6

To make sure our findings are robust and replicable, we conducted Experiment 6 as a replication of Experiment 5 whose sample size was determined by a simulation-based power analysis of Experiment 5.

4.5.1 Participants:

110 and 106 native English speakers were recruited via Prolific in Experiment 6a and 6b respectively.

4.5.2 Design and Materials:

The design and materials are the same as Experiment 5, except that participants read both simple transitives and clefts word-by-word in Experiment 6. Experiment 6a and 6b tested simple transitives and clefts, respectively.

4.5.3 Results:

Data from subjects who were not native speakers of American English or who did not answer all the comprehension questions in the fillers with at least 85% accuracy were excluded. Responses from 103 participants in Experiment 6a and 101 participants in Experiment 6b were analyzed.

A. Comprehension question task

We first analyzed data for simple transitives and clefts separately. Deploying the same models used in Experiment 5 - mixed-effects logistic regressions with word order (SVO vs. OSV or subject vs. object clefts) and plausibility (plausible vs. implausible) as predictors, as well as by-subject and by-item random intercepts. The results replicated Experiment 5 – in both datasets, (i) low-frequency sentences were more likely to be interpreted non-literally compared to high-frequency sentences (βs > 1.5, zs>3.9, ps<0.001); (ii) there were also significantly more non-literal responses for implausible sentences than plausible sentences (βs > 3.5, zs > 4.3, ps<0.001) as in Figs. 12 & 13.

To distinguish the construction-based hypothesis and the linear string-based hypothesis, we analyzed responses of the implausible condition for both simple transitives and clefts, using the same model in Experiment 5 - frequency (high vs. low), construction (simple transitives vs. clefts), and their interaction were entered as the predictors, as well as by-subject and by-item random intercepts. Replicating the results of Experiment 5, (i) a significant interaction between frequency and construction was observed (β=-2.98, z=-3.94, p<0.001), supporting the construction-based hypothesis; (ii) low frequency structures (OSV & clefts) were more likely to be interpreted non-literally (β=-3.43, z=-8.49, p<0.001), supporting our hypothesis about the structural prior.

In sum, the results of the comprehension task in Experiment 6 replicated our findings in Experiment 5.

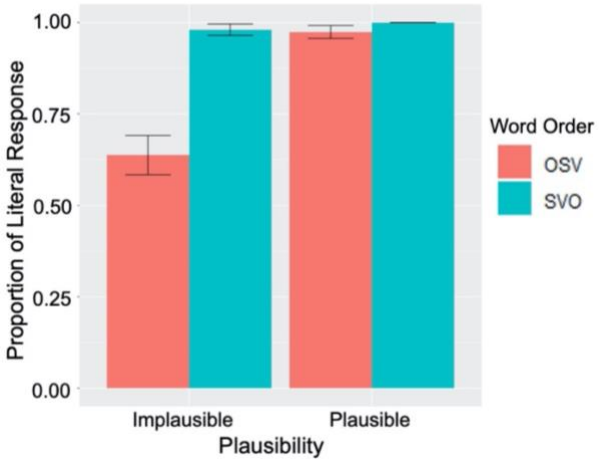


Fig.12. Proportion of choosing literal interpretation by word order (SVO vs. OSV) and plausibility (95% CI) in Experiment 6a.

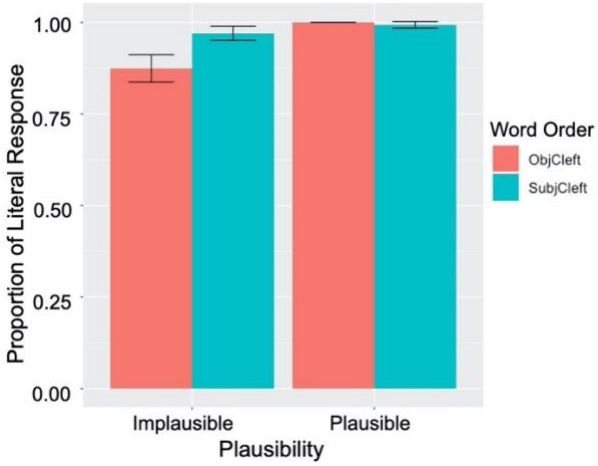


Fig.13. Proportion of choosing literal interpretation by word order (subject vs. object clefts) and plausibility (95% CI) in Experiment 6b.

B. Retyping task

Consistent with the results of Experiment 5, almost all participants accurately retyped the exact form of the input sentence, including those who provided non-literal responses to the comprehension questions, as in Figs. 14 & 15, supporting our hypothetical speaker's channel hypothesis.

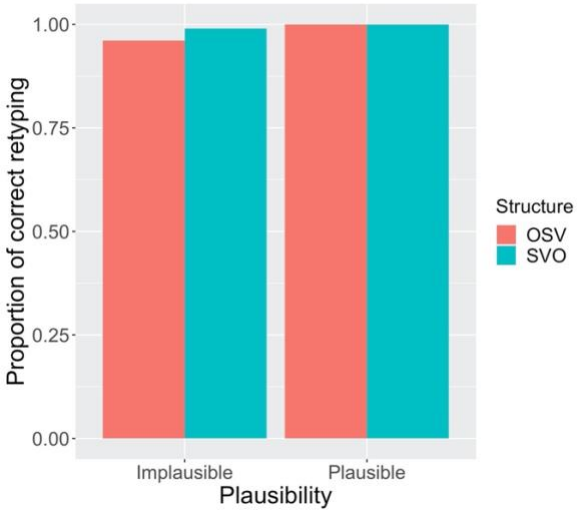


Fig.14. Proportion of correct retyping by word order (SVO vs. OSV) and plausibility (95% CI) in Experiment 6a.

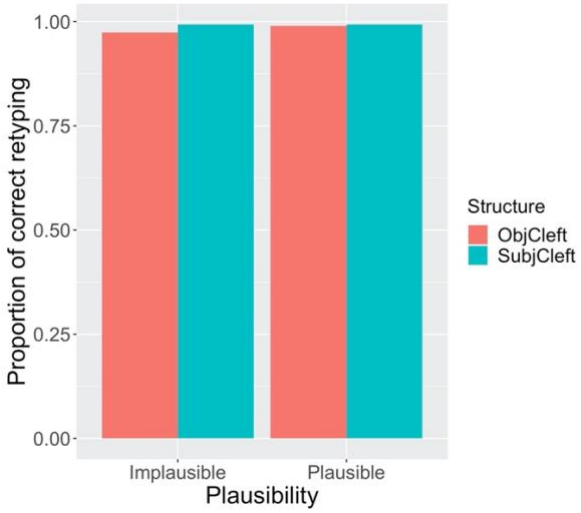


Fig.15. Proportion of correct retyping by structure (subject vs. object clefts) and plausibility (95% CI) in Experiment 6b.

4.5.4 Discussion

Experiment 6 replicated all the three major findings in Experiment 5: (i) there were significantly more non-literal interpretations for low-frequency structures than for high-frequency structures in both simple transitives and clefts, confirming the impact of structural frequency on noisy-channel processing ; (ii) the amount of non-literal responses was in proportion to the relative frequency contrast between the input and its neighborhood candidates, supporting the construction-based noisy-channel model; (iii) there was almost no erroneous retyping for responses of both literal and non-literal interpretations – comprehenders are fully aware of the input sentence when drawing inference, which align the rational noisy-channel inference.

Thus, the results in Experiment 6 provide exceptionally robust evidence for our proposed noisy-channel model with integration of syntactic information.

4.6 General Discussion

In Experiments 1-2, we found that Chinese allowed more flexible word order than English. Aligning the results of corpus search, SVO is be a more frequent/acceptable word order than OSV in both English and Chinese. Experiments 3-4 revealed that as predicted by our proposed noisy-channel model, OSV sentences were more likely to be interpreted non-literally compared to SVO sentences in both English and Mandarin. Experiment 5 demonstrated that the observation of more non-literal interpretations for low-frequency structures than for high-frequency structures was not only replicated in simple transitives but was also confirmed in clefts. More importantly, we found that the amount of inference was in proportion to the relative frequency contrast between the input and its neighborhood candidates, which supported the construction-based hypothesis, not the linear string-based hypothesis. We also observed

that there was almost no incorrect retyping for almost all the responses for both simple transitives and clefts, including the non-literal interpretations. That indicated that at comprehenders' non-literal responses did not come from misperception of the input sentence, and the non-literal interpretations were more likely to come from the comprehenders' inference about the speaker's intended utterance, supporting the hypothetical speaker's channel hypothesis. Experiment 6 replicated all the findings in Experiment 5.

Overall, this work has made three major contributions to the field of cognitive science and linguistics: (a) Comprehenders' noisy-channel processing is sensitive to syntactic information, in addition to meaning, in both English and Chinese. (b) As for what 'grain sizes' distributional syntactic information is stored by language users, this project shows that comprehenders track the input sentence at the level of construction, rather than the linear string of content words. (c) This project defines the two possible sources of the non-literal interpretation – rational inference vs. misperception. Through a novel experimental design, we distinguish these two hypotheses and show that the non-literal interpretations come from comprehenders' rational inferenc of the intended utterance, rather than misperception.

This project focuses on English and Mandarin where there is little to no case markings. From a cross-linguistic perspective, it's likely that the syntax prior in the noisy-channel model will not only encode structural frequency but also case markings (c.f. MacWhinney, 2021), especially for languages with richer morphologies. For instance, Kurumada & Jaeger (2015) have demonstrated that to achieve robust information transmission, Japanese speakers are more likely to produce optional object case-markings when the intended message is less plausible. Future works are needed to explore this area.

Chapter 5: Conclusion

This thesis has been an examination of how various factors shape people's acceptability and interpretation of long-distance dependencies. In three papers, I showed that frequency-based processing accounts and discourse accounts fare better in capturing different types of long-distance dependencies. I have also begun to better characterize the structural prior and the source of non-literal interpretation in noisy-channel comprehension. Below, I discuss the implications of these findings for discourse-based accounts, for the sentence processing literature, and for the innateness and learnability of long-distance dependencies.

5.1 Implications for discourse-based accounts of long-distance dependencies

Discourse-based accounts capture sentence acceptability/grammaticality via factors beyond the form, such as or perspective-taking or salience of information in different constitutes of the sentence. More specifically, according to the discourse accounts, island effects are due to fronting non-focused/non-salient information in filler-gap constructions, and the reference of reflexives is largely determined by the logophoric status of the antecedent (e.g., Goldberg, 2016; Charnavel, 2019). My reflexive project (paper 2) and many experimental works about the island phenomena (e.g., Abeille et al., 2020; Chaves & Putnam, 2020) have demonstrated that the discourse accounts can offer independently driven and testable explanations for a wide range of long-distance dependencies which were unexplained by a pure structural approach.

My island project (paper 1) and my reflexive project (paper 2) reveal that a future direction for the discourse accounts is to have better measurements of discourse/information salience of various constituents in the sentence. Though notions such as focus have theoretical definitions, the

corresponding empirical measurements have not been fully figured out. For instance, it still remains unclear whether the negation test adopted by Ambridge and Goldberg (2008) is a good measurement for backgroundedness (Liu et al., 2021). And the fact that we find no evidence for the discourse account in paper 1 might be partially due to lack of accurate measurement of backgroundedness. Similarly, paper 2 shows that the logophoric status of an antecedent is more than being absolutely logophoric or non-logophoric, as reflected by the gradience in sentence acceptability. A more fine-grained theoretical framework with better empirical measurements reflecting the gradience of discourse salience can be developed in future works.

5.2 Implications for sentence processing

Ample works have been done to investigate the structural forms that language users store to learn and process language. Though previous works have demonstrated that comprehenders are sensitive to prior linguistic experience in the sense of probabilistic context-free grammar rules, where the probabilities of all the constituents in a sentence are counted (e.g., surprisal in Hale (2001) and Levy (2008)), it is not yet known at what 'grain sizes' distributional syntactic/lexical information is stored by language users (e.g., Mitchell et al., 1995). For instance, there are various grain sizes that language users could be using to store object-clefts – they might keep track of *it-is*-NP-RC sequences, or more coarse-grained *it-is*-NP-Modifier sequences, or other possibilities, both more narrow or more general. At the more general level, people might alternatively pool the statistics for object-cleft and topicalization, keeping track of both constructions in the form N-N-V.

For sentence comprehension, our findings of cross-construction variations in paper 2 suggest that comprehenders track frequency at the level of construction (clefts vs. simple transitives) rather than

linear string of content words (NVN vs. NNV). An input sentence can only be recovered to its neighborhood candidates which are sufficiently similar to it. For instance, an object cleft might be recovered to a subject cleft, but not to a dative structure or an OSV structure.

Another vital point about sentence processing addressed by this dissertation is that non-literal interpretations in the literature might come from different sources. In paper 3, our findings suggest that comprehenders' non-literal interpretations for simple transitives and clefts come from comprehenders' rational inference of the speaker's intended message rather than misperception of the input. But mishearing and misreading do exist in daily communication, which might result in non-literal interpretations in other phenomena.

5.3 Implications for innateness and learnability

In paper 1, our finding that the acceptability of wh-questions is highly correlated with verb-frame frequency suggests that the unacceptability of certain filler-gap constructions is modulated by exposure, and is therefore learnable. That suggests although direct negative evidence is missing especially for such complex structures, children may draw statistical inferences from the input and regard the absence of a certain input (e.g., a type of extraction) as evidence of its oddness (rendering it unacceptable) (cf. Hsu & Griffiths, 2016; Kidd, Lieven & Tomasello, 2010; Navarro, Dry & Lee, 2012; Voorspoels, Perfors, Ransom & Storms, 2015; Xu & Tenenbaum, 2007).

Though we did not find support for the island effects or unacceptable anaphoric dependencies being unlearnable, this work does not deny the importance of syntactic structure in language processing and learning. Indeed, by considering alternatives to covert structures that are not supported by independent

empirical evidence, we may in fact reach a more efficient and simpler syntactic framework (c.f., Culicover & Jackendoff, 2005).

Appendix A: Four Analyses of Experiment 1

Here we present four analyses relevant to Experiment 1:

(I) An ordinal regression analysis applied to our collected data, to test the discourse BCI account.

(II) A Bayes factor analysis to evaluate the evidence for and against the presence of the BCI effect.

(III) Model comparison to assess whether verb-frame frequency offers a better explanation for the observed data than the BCI account.

(IV) A re-analysis of data from Ambridge and Goldberg (2008) using ordinal regression.

We thank Ben Ambridge for making the original data in A&G (2008) publicly available.

I. Application of ordinal regression to our collected data for the BCI account

We fit two ordinal logit regressions on our data of Experiment 1 based on the BCI account, using the *ordinal* package in R. In both of these two models, we entered *sentence type* (declarative vs. wh-question), *mean negation scores*, and their interaction as the predictors, as in Table 1(a&b) below. One model (a) was fit on all the 24 tested verbs, and another (b) was applied to 23 verbs, excluding the verb 'know', as this verb is potentially pragmatically special within wh-questions. The two models were fit with the maximum random effect structure which allowed the models to converge. The model fit on 24 verbs (a) contained random intercepts for *subjects* and *verbs* as well as by-subject slopes for the effects of *sentence type*, *negation scores*, and their *interaction* and by-verb *sentence type* slopes. The other model with 23 verbs (b) has the same random effect structure as (a), except that the random slope of the *interaction* between sentence type and negation scores was removed to facilitate convergence.

The BCI account predicts a significant interaction between sentence type and mean negation scores. Model (a) fit on all the 24 verbs showed that sentence type is a significant predictor for acceptability, but no significant interaction was found (β=0.32, Z=0.195, p=0.0512). Model (b) with 23 verbs (excluding 'know') showed a smaller effect for the interaction (β=0.16, Z=1.05, p=0.29), suggesting the non-significant marginal interaction effect might be in part driven by 'know'.

The results of the two models in Table 1(a&b) are consistent with our previous findings in Experiment 1.

Table 1: Ordinal regression for the BCI account with the interaction effect:

Model: Rating~sentence_type*mean_neg			
	β	z value	p value
a. Model fit with all the 24 tested verbs			
sentence_type	-2.25729	-4.418	9.95e-06 ***
mean_neg	0.02984	0.154	0.8779
sentence_type:mean_neg	**0.32200**	**1.950**	**0.0512**
b. Model fit with 23 tested verbs, excluding 'know'			
sentence_type	-1.7260	-3.578	0.000347 ***
mean_neg	0.0226	0.107	0.915137
sentence_type:mean_neg	**0.1601**	**1.048**	**0.294752**

II. A Bayes factor analysis of the interaction effect between sentence type and negation scores

We fit another ordinal model (Table 2) to our collected data in Experiment 1 for all 24 tested verbs, entering *sentence type* (declarative vs. wh-question) and *mean negation scores* as the predictors but without their interaction. The model was fit with the maximum random effect structure which included random intercepts for *subjects* and *verbs* as well as by-subject slopes for the effects of *sentence type* and *negation scores*, and by-verb *sentence type* slopes. We then compared the Bayesian Information Criterion (BIC) of the two models with (Table 1a) and without (Table 2) the interaction between sentence type and negation scores, as in (Table 3a). We found that the model without this interaction has a 28.96 smaller BIC than the one with the interaction. A 28.96 difference in BIC is generally considered as strong evidence favoring the model without the interaction (Raftery, 1995).

We further calculated the Bayes factor for this interaction effect based on the BIC estimates of these two models in (Table 3b). Different from p-values, which only provide evidence for how unlikely the data are under the null hypothesis, Bayes factor allows us to compare the likelihood of the data under the alternative hypothesis with the likelihood of the data under the null hypothesis (BF_{10}). The higher a Bayes factor (BF_{10}), the more evidence in support of the alternative hypothesis. The lower a Bayes factor (BF_{10}), the more evidence for the null hypothesis. The Bayes factor for the interaction between sentence type and negation scores is below 0.0001, which is strong evidence for H0, no interaction effect (Schonbrodt & Wagenmakers, 2018).

The results of analyses (I) & (II) are consistent with our reported results from the ordinal and logistic regressions in the main text, suggesting no robust interaction effect between sentence type and negation scores.

Table 2: Ordinal regression without interaction between sentence type and negation scores

Model: Rating ~ sentence_type+mean_neg (24 verbs)			
	β	z value	p value
sentence_type	-1.34609	-6.86	6.87e-12 ***
mean_neg	-0.07645	-0.39	0.697

Table 3

a.BIC of the two ordinal regressions fit with and without the interaction effect		
Model (24 verbs)	df	BIC
Rating ~ sentence_type+mean_neg	15	**21385.03**
Rating ~ sentence_type*mean_neg	20	**21413.99**
b. Bayes Factor for the interaction effect: exp((21385.03 - 21413.99)/2) = **0.0000005**		

III. Model comparison for the discourse BCI and our frequency accounts

We conducted a model comparison between models fit according to the discourse BCI (Table 1a) and our verb-frame frequency (Table 4) accounts. Ordinal regression in (Table 4) was fit with two predictors - *sentence type* and *log-transformed verb-frame frequency*, with the maximum random effect structure, containing random intercepts for *subjects* and *verbs* as well as by-subject slopes for the effects of *sentence type* and *log-transformed frequency*, and by-verb *sentence type* slopes. We did not include an interaction between these two predictors, because the verb-frame frequency account predicts no interaction, and there was no evidence for such an interaction effect when it was included in the model ($p>0.08$, as reported above in the results of Experiment 1). Model comparison in (Table 5) shows that the frequency-based model has a 428.97 lower BIC, which suggests that the verb-frame frequency account offers a more parsimonious explanation for the observed data.

Table 4: Ordinal regression for our verb-frame frequency account:

Model: Rating ~ sentence_type+log_fre (24 verbs)			
	β	z value	p value
sentence_type	-1.401	-6.838	8.03e-12 ***
log_fre	0.494	5.520	3.39e-08 ***

Table 5: Model comparison for the discourse BCI and our frequency accounts

Model (24 verbs)	df	BIC
Rating ~ sentence_type*mean_neg	20	**21413.99**
Rating ~ sentence_type+log_fre	15	**20985.02**

IV. Ordinal regression analysis for data from Ambridge and Goldberg (2008)

We fit two ordinal logit regressions on the dataset of A&G (2008), using the *ordinal* package in R. In both of these two models, we entered *sentence type* (declarative vs. wh-question), *mean negation scores*, and their interaction as the predictors. The models were fit with the maximum random effect structure. One model (Table 6a) was fit on all the 12 tested verbs, and another (Table 6b) was applied to 11 verbs, excluding the verb 'know'. Results of both models showed that both sentence type and the interaction between sentence type and negation scores are significant predictors of acceptability ratings. These results from ordinal regressions are consistent with the results reported in the original paper A&G (2008).

Table 6: Ordinal regression to the original data in A&G(2008)

Model: Rating~sentence_type*mean_neg			
	β	z value	p value
a. Model fit with all the 12 tested verbs			
sentence_type	-5.4102	-7.824	5.10e-15 ***
mean_neg	0.1215	0.448	0.654
sentence_type:mean_neg	1.0312	4.411	1.03e-05 ***
b. Model fit with 11 tested verbs, excluding 'know'			
sentence_type	-4.4108	-6.870	6.41e-12 ***
mean_neg	0.2528	0.826	0.408916
sentence_type:mean_neg	0.7114	3.306	0.000946 ***

In addition to the analyses in (I) - (IV), Table 7 is a summary of three models fit on our collected data in Experiment 1: model in (a) is the frequency-based model with *sentence type* and *verb-frame frequency* as predictors; model (b) is the discourse-based model, including predictors of *sentence type*, *negation scores* and their *interaction*; model in (c) includes both discourse- and frequency- based factors as fix effects. All the three models were fit with maximal random effect structures. These models were summarized based on four dimensions - BIC, AIC (Akaike Information Criterion), Log-likelihood and degree of freedom.

The discourse-only model (b) has the highest BIC and lowest log-likelihood. Based on BIC, we favor the frequency-based model (a). Note that the log-likelihood of the model including both discourse and frequency factors (c) has the largest log-likelihood, suggesting the discourse factor (interaction effect between sentence type and negation score) helps to explain some of the variance in the observed data, though the captured variance might be relatively small so that it's hard to find robust evidence for it.

Table 7: Summary of three models

Model (ST = sentence type)	BIC	AIC	Log-likelihood	df
(a)Rating ~ ST+fre	20985	20875	-10423	15
(b)Rating ~ ST*neg	21414	21268	-10614	20
(c)Rating ~ ST*neg + fre	21047	20857	-10402	26

Appendix B: Experiment 4 - A 5-point Likert Scale Version of Experiment 3

We also ran a 5-point Likert scale version of Experiment 3 with the same materials and design. We applied mixed effects ordinal logit regression in the *ordinal* package in R to the data (Table 8). The results were similar to Expt 3. Sentence type (declaratives vs. clefts) and frequency were significant predictors of acceptability, while no reliable interaction was found.

Table 8: Ordinal model

Model: Rating ~ sentence_type (decl vs. cleft)*log_fre			
	β	z value	p value
sentence_type	-4.40696	-10.792	< 2e-16 ***
log_fre	0.65663	6.681	2.37e-11 ***
sentence_type:log_fre	0.10389	0.836	0.403

Different from the results of the ordinal regression in Table 8, a linear model with the same predictors (Table 9) applied to the same set of data showed a significant interaction between sentence type and frequency. These results are consistent with Liddell & Kruschke (2018) that application of linear regression on ordinal data could lead to false positives or false negatives.

Table 9: Linear model

Model: Rating ~ sentence_type (decl vs. cleft)*log_fre			
	β	t value	p value
sentence_type	-1.48789	-11.648	< 2e-16 ***
log_fre	0.21134	7.517	3.5e-10 ***
sentence_type:log_fre	0.09735	2.859	0.00625 **

Appendix C: Full Table of Results

Below are the full table of results of all the regressions reported in the main text of paper 1.

Experiment 1

Table 10: Ordinal regression for 5-point Likert scale ratings

Model: Rating ~ sentence_type (decl vs. wh-q)*log_fre			
	β	z value	p value
sentence_type	-1.4022	-7.038	1.96e-12 ***
log_fre	0.5012	5.889	3.88e-09 ***
sentence_type:log_fre	0.1886	1.712	0.0869

Table 11: Logistic regression (transformation of rating 1-2 to 0 and rating 3-5 to 1)

Model: Rating ~ sentence_type (decl vs. wh-q)*log_fre			
	β	z value	p value
sentence_type	-2.05054	-6.683	2.35e-11 ***
log_fre	0.44709	3.845	0.00012 ***
sentence_type:log_fre	-0.08663	-0.440	0.65991

Experiment 2

Table 12: Logistic regression for binary acceptability ratings

Model: Rating ~ sentence_type (decl vs. wh-q)*log_fre			
	β	z value	p value
log_fre	0.5888	3.947	7.92e-05 ***
Sentenece_type	-2.4501	-7.877	3.35e-15 ***
sentence_type:log_fre	-0.1791	-0.811	0.417

Experiment 3

Table 13: Logistic regression for binary acceptability ratings

Model: Rating ~ sentence_type (decl vs. cleft)*log_fre			
	β	z value	p value
sentence_type	-10.7127	-4.941	7.76e-07 ***
log_fre	1.2448	2.394	0.0167 *
sentence_type:log_fre	-0.8715	-0.841	0.4001

References

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction from subject: differences in acceptability depend on the discourse function of the construction. *Cognition 204, 104293.*

Ambridge, B., & Goldberg, A. E. (2008). The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, *19*(3). https://doi.org/10.1515/COGL.2008.014

Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2014). Child language acquisition: Why universal grammar doesn't help. *Language 90*(3). 53–90. https://doi.org/10.1353/lan.2014.0051.

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. https://doi.org/10.1016/j.jml.2007.12.005

Baltin, M. R. (1982). A landing site theory of movement rules. *Linguistic Inquiry*, *13*, 1–38.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Available online at http://lme4.r-forge.r-project.org/book/.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition*, 11, 245–299.

Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the thirty-fourth annual conference of the cognitive science society* (p. 1320-1325).

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190. http://doi.org/10.1016/j.jml.2009.04.003

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as Contexts in Conversation. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 62, pp. 59–99). Academic Press. http://doi.org/10.1016/bs.plm.2014.09.003

Chafe, W. L. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed), *Coherence and grounding in discourse* (pp. 5-25). Amsterdam: Benjamins.

Charnavel, I. (2014). Perspectives on Binding and Exemption. Talk given at Massachusetts Institute of Technology, United States.

Charnavel, I. (2019). *Locality and Logophoricity: A Theory of Exempt Anaphora*. Oxford: Oxford University Press.

Charnavel, I., & Dominique, S. (2016). Anaphor Binding: What French Inanimate Anaphors Show. *Linguistic Inquiry, 47* (1), 35–87.

Charnavel, I, & Huang, Y. (2018). Inanimate ziji and Condition A in Mandarin. In *35th West Coast Conference on Formal Linguistics* (pp.132–141). Somerville, MA: Cascadilla Proceedings Project.

Charnavel, I., Huang, J. C.-T., Cole, P., & Hermon, G. (2017). Long-distance anaphora: syntax and discourse. In M. Everaert & H. C. V. Riemsdijk (Eds.), *The Wiley Blackwell companion to syntax* (2nd ed.). Hoboken, NJ: John Wiley & Sons Inc.

Chierchia, G. (1989). Anaphora and Attitudes *de Se*. In R. Bartsch, J. V. Benthem & P. V. E. Boas (Eds.), *Semantics and Contextual Expression*, 1–32. Dordrecht: Foris.

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

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes, 42*, 205–238. doi: 10.1207/s15326950dp4202_6

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology, 69*, 817–828. doi: 10.1080/17470218.2015.1134603

Clements, G. (1975). The Logophoric Pronoun in Ewe: Its Role in Discourse. *Journal of West African Languages, 10*, 141–177.

Cole, P., & Hermon, G. (1998a). Long Distance Reflexives in Singapore Malay: An Apparent Typological Anomaly. *Linguistic Typology, 2*, 57–77.

Cole, P., & Hermon, G. (1998b). VP Ellipsis and Malay Reflexives. In *Proceedings of the 13th Annual Conference of the Israeli Association for Theoretical Linguistics*, edited by Adam Zachary Wyner, 39–54.

Cole, P., & Wang, C. (1996). Antecedents and Blockers of Long Distance Reflexives. *Linguistic Inquiry, 27*, 357–390.

Cole, P., Hermon, G., & Huang, C.-T. J. (2001). *Long Distance Reflexives*. New York, NY: Academic Press.

Cole, P., Hermon, G., & Sung, L. (1990). Principles and Parameters of Long-Distance Reflexives. *Linguistics Inquiry, 21*, 1-22

Cole, P., Hermon, G., & Sung, L. (1993). Feature Percolation and Mandarin Reflexives. *Journal of East Asian Linguistics, 2*, 91–118.

Chomsky, N. (1964). Current issues in linguistic theory. In J. A. Fodor and J. J. Katz (Eds.), *The structure of language: readings in the philosophy of language* (pp. 50-118). Englewood Cliffs, NJ: Prentice Hall.

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

Chomsky, N. (1977). On wh-movement. In P. Culicover, T. Wasow, A. Akmajian (Eds.), *Formal syntax* (pp. 71-132), New York: Academic Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Berlin: Mouton de Gruyter.

Chomsky, N. (1986a). *Barriers*. Cambridge: MIT Press.

Chomsky, N. (1986b).*Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Oxford University Press.

Dąbrowska, E. (2008). Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics*, *19*(3). https://doi.org/10.1515/COGL.2008.015

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*(1), 1–23. https://doi.org/10.1515/tlir.2010.001

De Cuba, C. (2018). Manner-of-speaking that-complements as close apposition structures. *Proceedings of the Linguistic Society of America*, *3*(1), 32. https://doi.org/10.3765/plsa.v3i1.4320

Deane, P. (1991). Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics*, *2*(1), 1–64. https://doi.org/10.1515/cogl.1991.2.1.1

Dillon, B., Chow, W. Y., and Xiang, M. (2016). The relationship between anaphor features and antecedent retrieval: comparing Mandarin ziji and ta-ziji. *Frontiers in Psychology, 6,* 1966.

Erteschik-Shir, N. (1973). *On the nature of island constraints*. PhD dissertation, MIT.

Erteschik-Shir, N. (1979). Discourse Constraints on Dative Movement. In T. Givon (Ed.), *Discourse and Syntax* (pp. 441–467). BRILL. https://doi.org/10.1163/9789004368897_019

Erteschik-Shir, N. (1998). *Dynamics of focus structure*. Cambridge University Press.

Erteschik-Shir, N. (2007). *Information structure: The syntax-discourse interface*. Oxford University Press.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47,* 164-203.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: implications for models of reanalysis. *Journal of Psycholinguistic Research, 30*, 3-20.

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

Fine, A., Jaeger, F., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE , 8* , e77661.

Geisler, W. S., & Diehl, R. L. (2003). A bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science, 27*, 379-402.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*, 8051-6.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't Underestimate the Benefits of Being Misunderstood. *Psychological Science, 1-10*. http://doi.org/10.1177/0956797617690277

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

Gibson, E., Piantadosi, S. T., Ichinco, D., & Fedorenko, E. (2012). Evaluating structural overlap across constructions: inter-subject analysis of covariation. In 86th Annual Meeting of the LSA, Portland, OR.

Giorgi, A. (1984). Towards a Theory of Long-Distance Anaphors: A GB Approach. *Linguistic Review*, *3*, 307–359.

Goldberg, A. (2013). Backgrounded constituents cannot be "extracted", in J. Sprouse, & H.Norbert (Eds.), *Experimental syntax and island effects* (pp. 221-238.). Cambridge: Cambridge University Press.

Goldberg, A. E. (2016). Subtle implicit language facts emerge from the functions of constructions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02019

Hagège, C. (1974). Les pronoms logophoriques. *Bulletin de la Société de Linguistique de Paris, 69*, 287–310.

Hale, J. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of NAACL* (Vol.2, pp. 159–166).

Hale, J. (2003). Grammar, uncertainty and sentence processing. PhD dissertation, John Hopkins University.

Harrington S., C., James, A. & Watson, D. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415.

Hsu, A., & Griffiths, T. L. (2016). Sampling Assumptions Affect Use of Indirect Negative Evidence in Language Learning. *PLOS ONE, 11*(6), e0156597. https://doi.org/10.1371/journal.pone.0156597

Huang, C-T. J. (1982). *Logical relations in Chinese and the theory of grammar*. PhD dissertation, MIT.

Huang, C.-T. J., and Tang, C.-C. J. (1991). "The local nature of the long-distance reflexive in Chinese," in *Long-Distance Anaphora*, eds J. Koster and E. Reuland (Cambridge: Cambridge University Press), 263–282.

Huang, C.-T. J., Huang, Y.-H., Teng, T.-H., and Tiedeman, R. (1984). "Reflexives in Chinese and the teaching of Chinese," in *Proceedings of the First World Conference on Chinese Language* (Taipei), 205–215.

Huang, C.-T. J., and Liu, L. (2001). "Logophoricity, attitudes, and *ziji* at the interface," in *Long Distance Reflexives, Syntax and Semantics*, Vol. 33, eds P. Cole, C.-T. J. Huang, and G. Hermon (New York, NY: Academic Press), 141–195.

Hoekstra, T., & Kooij, J. G. (1988). The innateness hypothesis. In John A. Hawkins (ed.), *Explaining language universals*, 31-55. Oxford, UK: Blackwell.

Ionin, T., Ko, H., & Wexler, K. (2004). Article Semantics in L2 Acquisition: The Role of Specificity. *Source: Language Acquisition*, *12*(1), 3–69.

Jaeger, F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology, 61*, 23-62.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. MIT Press.

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology, 124*, 101359.

Kidd, E., Lieven, E. V. M., & Tomasello, M. (2010). Lexical frequency and exemplar-based learning effects in language acquisition: Evidence from sentential complements. *Language Sciences*, *32*(1), 132–142. https://doi.org/10.1016/j.langsci.2009.05.002

Kiparsky, P., and Kiparsky, C. (1971). Fact. In M. Bierwisch and K. Heidolph (Eds.), *Progress in Linguistics* (pp. 143–173). The Hague: Mouton.

Kleinschmidt, D., & Jaeger, F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148-203.

Kothari, A. (2008). Frequency-based expectations and context influence bridge quality. In M. Grosvald, & D. Soares (Eds.), *Proceedings of WECOL 2008*. UC Davis Department of Linguistics; 2008. http://www.stanford.edu/~anubha/publications.html.

Kurumada, C., Brown, M., & Tanenhaus, M. (in press). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-017-1332-6

Kush, D., Lohndal, T. and Sprouse, J. (2019). On the Island Sensitivity of Topicalization in Norwegian: An experimental investigation. *Language, 95*(3), 393-420.

Kuno, S. (1972). Pronominalization, Reflexivization and Direct Discourse. *Linguistic Inquiry, 3*, 161–195.

Kuno, S. (1978). *Danwa no bunpô* [*Grammar of Discourse*]. Tokyo: Taishukan.

Kuno, S. (1987). *Functional Syntax: Anaphora, Discourse and Empathy*. Chicago, IL: University of Chicago Press.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge University Press.

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

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 21086-90.

Li, P., & MacWhinney, B. (2013). Competition model. In C. A. Chapelle (Ed.), *The Encyclopaedia of Applied Linguistics* (pp.1–5). Blackwell Publishing Ltd. doi:10.1002/ 9781405198431.wbeal0168

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Liu, Y. (2016). *Chinese Zi: Linking Reflexivization and Binding*. MA thesis, Utrecht University, Utrecht, the Netherlands.

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2019). Verb frequency explains the unacceptability of factive and manner-of-speaking islands in English. In Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp.685–691). Montreal, QC: Cognitive Science Society.

Liu, Y.*, Winckel, E.*, Abeillé, A., Hemforth, B., & Gibson, E. (2021). *Structural, functional and processing perspectives on linguistic island effects.* Manuscript submitted for publication.

Kishida, M. (2012). On the argument structure of Zi-verbs in Japanese: reply to Tsujimura and Aikawa (1999). *Journal of East Asian Linguistics*, *21*, 197-218.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

MacWhinney, B. (1977). Starting Points. *Language, 53*(1), 152-168. https://doi.org/10.2307/413059

MacWhinney, B. (1992). Transfer and Competition in Second Language Learning. *Advances in Psychology, 83*, 371–390. https://doi.org/10.1016/S0166-4115(08)61506-X

MacWhinney, B. (2010). Computational models of child language learning. *Journal of Child Language, 37*, 477-485.

MacWhinney, B. (2005). Extending the Competition Model. *International Journal of Bilingualism, 9*, 69-84. https://psyling.talkbank.org/years/2005/ijb.pdf

MacWhinney, B. (2021). The Competition Model: Past and Future. In J. Gervain (Ed.), A life in cognition (pp. 3-16). Springer Nature.

Michaelis, L. & Hartwell, F. (2007). Lexical subjects and the conflation strategy. In N. Hedberg & R. Zacharski (Eds.), *Topics in the grammar–pragmatics interface: Papers in honour of Jeanette K. Gundel* (19-48). Amsterdam: John Benjamins.

Mitchell, D.C., Cuetos, F., Corley, M.M.B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research, 24*, 469-488.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling Assumptions in Inductive Generalization. *Cognitive Science, 36*(2), 187–223. https://doi.org/10.1111/j.1551-6709.2011.01212.x

Newmeyer, F. J. (1991). Functional explanation in linguistics and the origins of language. *Language & Communication, 11*(1–2). 3–28. https://doi.org/10.1016/0271-5309(91)90011-J.

Nicoladis, E. (2018). Cross-linguistic transfer in adjective–noun strings by preschool bilingual children. *Bilingualism: Language and Cognition, 9*(1), 15–32. https://doi.org/10.1017/S136672890500235X

Nishigauchi, T. (2014). Reflexive Binding: Awareness and Empathy from a Syntactic Point of View. *Journal of East Asian Linguistics, 23*, 157–206.

Pan, H. (1997). *Constraints on Reflexivization in Mandarin Chinese*. New York, NY: Garland.

Phillips, Colin. (2006). The real-time status of island phenomena. *Language, 82*(4), 795–823. https://doi.org/10.1353/lan.2006.0217

Pica, P. (1987). On the nature of the reflexivization cycle. In Proceedings of the North Eastern Linguistic Society 17, ed. J. McDonough and B. Plunkett, 483–499

Pica, Pierre. (1987). On the Nature of the Reflexivization Cycle. *Proceedings of the North Eastern Linguistic Society*, 17, 483–499.

Pollard, C., & Ivan, A. (1992). Anaphors and the Scope of Binding Theory. *Linguistic Inquiry, 23*, 261–303.

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proceedings of the 38th annual meeting of the cognitive science society* (p. 378-383). Poster presentation.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry, 24*, 657–720

Reuland, E. (2011). *Anaphora and language design. Linguistic inquiry monographs.* Cambridge, MA: MIT Press.

Reuland, E. (2018). Reflexives and Reflexivity. *Annual Review of Linguistics, 4*, 81-107

Richter, S., & Chaves, R. (2020). Investigating the role of verb frequency in factive and manner-of speaking islands. In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (pp.1771-1777), Toronto, ON: Cognitive Science Society.

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: MIT Press.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of memory and language*, *57*(3), 348-379.

Ross, J. R. (1967). *Constraints on variables in syntax*. PhD dissertation, MIT. http://hdl.handle.net/1721.1/15166

Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-schmidt, S. (2017). Verb Biases Are Shaped Through Lifelong Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 781–794.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, 181, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018

Sag, I. A. (2010). English filler-gap constructions. *Language*, *86*(3), 486–545. https://doi.org/10.1353/lan.2010.0002

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379-423.

Schönbrodt, F.D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schütze, C. T., Sprouse, J., & Caponigro, I. (2015). Challenges for a theory of islands: A broader perspective on Ambridge, Pine, and Lieven. *Language*, *91*(2), e31–e39. https://doi.org/10.1353/lan.2015.0014

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from flawed semantic processing. *Journal of Memory and Language, 69*, 104–120. doi: 10.1016/j.jml.2013.04.001

Sells, P. (1987). Aspects of Logophoricity. *Linguistic Inquiry, 18*, 445–479.

Shuai L., Gong T., Wu Y.-C. (2013). Who is who? Interpretation of multiple occurrences of the Chinese reflexive: evidence from real-time sentence processing. PLoS ONE. 8:e0073226.

Sloggett, S., & Dillon, B. (2018). Person blocking in reflexive processing: When 'I' matter more than 'them'. Paper presented at 31st Annual CUNY Sentence Processing Conference, University of California, Davis.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control , 7*(1), 1–22.

Sundaresan, S. (2012). *Context and (Co)Reference in the Syntax and Its Interfaces*. Ph.D. thesis, University of Tromsø and University of Stuttgart.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, *34*(1), 307–344. https://doi.org/10.1007/s11049-015-9286-8

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

Snyder, W. (1992). Wh-extraction and the lexical representation of verbs. Unpublished manuscript., MIT, Cambridge, MA.

Stoica, I. (2016). Island effects and complementizer omission: the view from manner of speaking verbs. In P. Petrar, & A. Precup (Eds.), *Constructions of Identity (VIII): Discourses in the English-Speaking World* (pp. 191-200). Cluj-Napoca, România: Presa Universitară Clujeană. http://www.editura.ubbcluj.ro/bd/ebooks/pdf/2036.pdf

Stowell, T. A. (1981). *Origins of phrase structure*. PhD Dissertation, MIT.

Tang, C.-C. J. (1989). Chinese reflexives. *Natural Language and Linguistic Theory*, 7, 93–121.

Tonhauser, J., Beaver, D. I., & Degen, J. (2018). How Projective is Projective Content? Gradience in Projectivity and At-issueness. *Journal of Semantics*, *35*(3), 495–542. https://doi.org/10.1093/jos/ffy007

Van Valin, R. D. Jr. (1998). The acquisition of wh-questions and the mechanisms oflanguage acquisition. In M. Tomasello (Eds.), *The new psychology of language: Cognitive and functional approaches to language structure*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Valin, R. D., & LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press.

Verhagen, A. (2005). *Constructions of intersubjectivity discourse, syntax, and cognition*. http://www.ebrary.com

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25. https://doi.org/10.1016/j.cogpsych.2015.07.001

Wong, S. (2017). Investigating Mandarin Chinese zi-V reflexive verbs. *Working paper of Utrecht Institute of Linguistics,* Utrecht University, Utrecht, the Netherlands.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245.

Xu, L. (1993). Long Distance Binding of *Ziji*. *Journal of Chinese Linguistics*, *21*, 123–141.

Xu, L. (1994). The Antecedent of *Ziji*. *Journal of Chinese Linguistics*, *22*, 115–137.

Yu, X.-F. W. (1992). Challenging Chinese Reflexive Data. *Linguistic Review*, *9*, 285–294.

Yu, X.-F. W. (1996). *A Study of Chinese Reflexives*. Ph.D. thesis, University of London, London, UK.

Zwicky, A. (1971). In a Manner of Speaking. *Linguistic Inquiry*, *2*(2), 223-233.

Zribi-Hertz, A. (1989). Anaphor Binding and Narrative Point of View. *Language*, *65*, 695–727.