A Verb-frame Frequency Account of Constraints on Long-Distance Dependencies in English

Yingtong Liu[1,2] (y_liu@g.harvard.edu), Rachel Ryskin[3] (rryskin@ucmerced.edu), Richard Futrell[4] (rfutrell@uci.edu) , Edward Gibson[2] (egibson@mit.edu)

[1] Department of Linguistics, Harvard University, Cambridge, MA 02138 USA
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139 USA
[3] Department of Cognitive and Information Sciences, University of California, Merced, CA 95343 USA
[4] Department of Language Science, University of California, Irvine, CA 92697 USA

Abstract

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 the present work, 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. Here we provide 3 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.

Keywords: Sentence processing; Frequency effects; Long-distance dependencies; Syntactic islands

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 __$_i$

[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).

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.

(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).

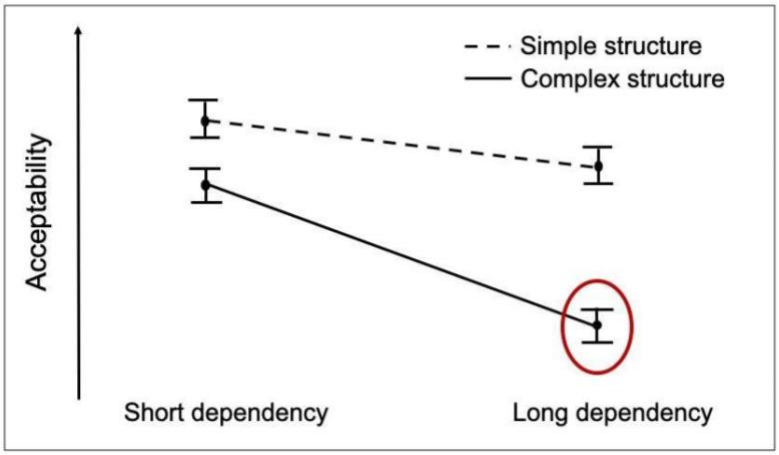


Figure 1. Illustration of a super-additive island effect, such that the complex, long dependency structure is rated least acceptable of the four conditions, and there is an interaction between dependency length (short vs. long) and complexity of the structures (complex vs. simple).

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

[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).

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 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.

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.

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.
~~→~~ Sentence 2: Mary didn't buy a car.
b. Sentence 1: I **didn't shout** that Mary bought a car.
?~~→~~ 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 $\_\__i$]?

[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.

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.

Syntactic Accounts

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 **[$_{\text{lightV}}$(made)][$_{\text{NP}}$ (a) [$_{\text{N}}$ 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. But a serious problem with this kind of account is that there are no independent reasons to propose these covert complex structures.

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

[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.').

(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.

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 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 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.

1.2 Predictions of the Four Theories on Factive and Manner-of-speaking Islands

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

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

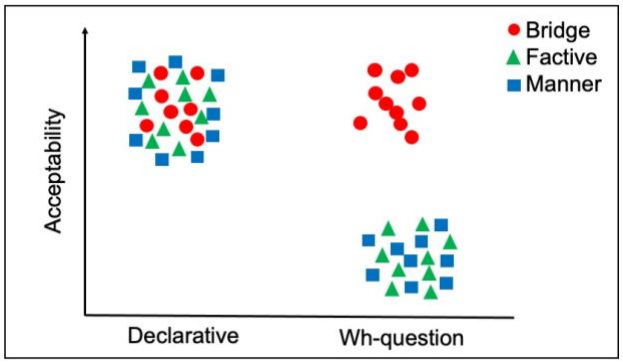
The syntactic accounts predict that all factive and manner-of-speaking wh-questions should be less acceptable than all the bridge ones due to categorically distinct covert structures which forbid extraction (e.g., Kiparsky & Kiparsky, 1971; Stowell, 1981; Snyder, 1992), as in Figure 2a.

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).

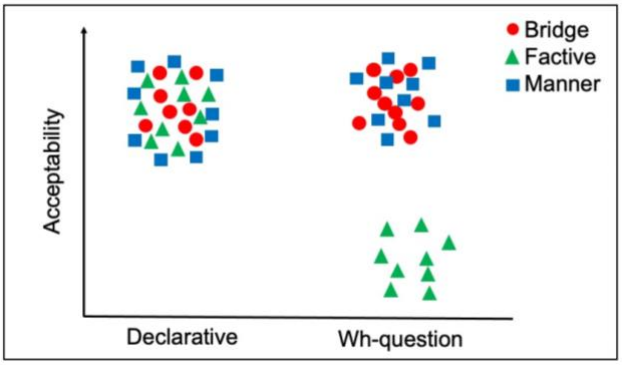
The BCI account (A&G 2008) predicts that the more backgrounded the sentence complement is, the less acceptable the wh-question. A&G (2008) measured wh-question acceptability by calculating the difference score between ratings of declaratives and the corresponding wh-questions -- higher difference scores 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.

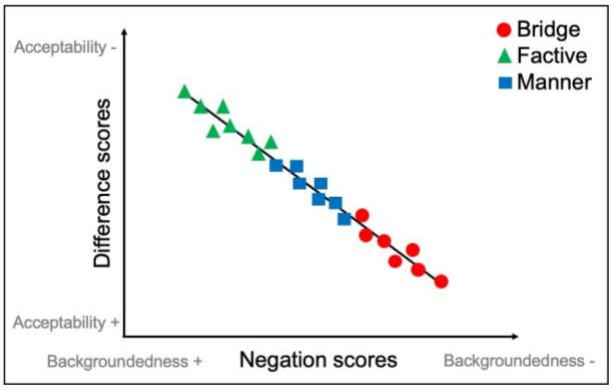
2a. Predictions of the syntactic accounts.



2b. Predictions of the semantic accounts.



2c. Predictions of the discourse BCI account.



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

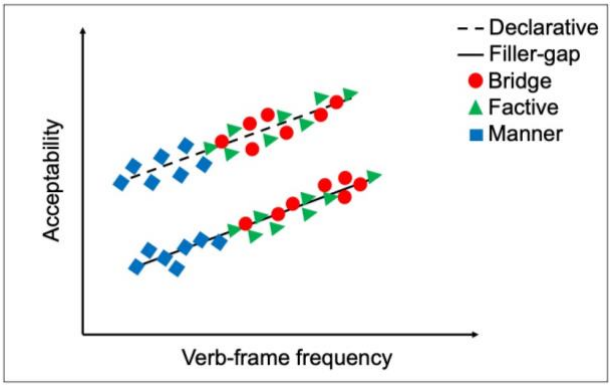


Figure 2. Predictions of the syntactic, semantic, discourse and frequency accounts. Each dot represents a word (conceptually). In Figures 2a, 2b and 2d, the y-axis is the raw rating. In Figure 2c, the y-axis denotes the difference scores between ratings of wh-question and declaratives (following Ambridge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

2.1 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 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.

2.3 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.

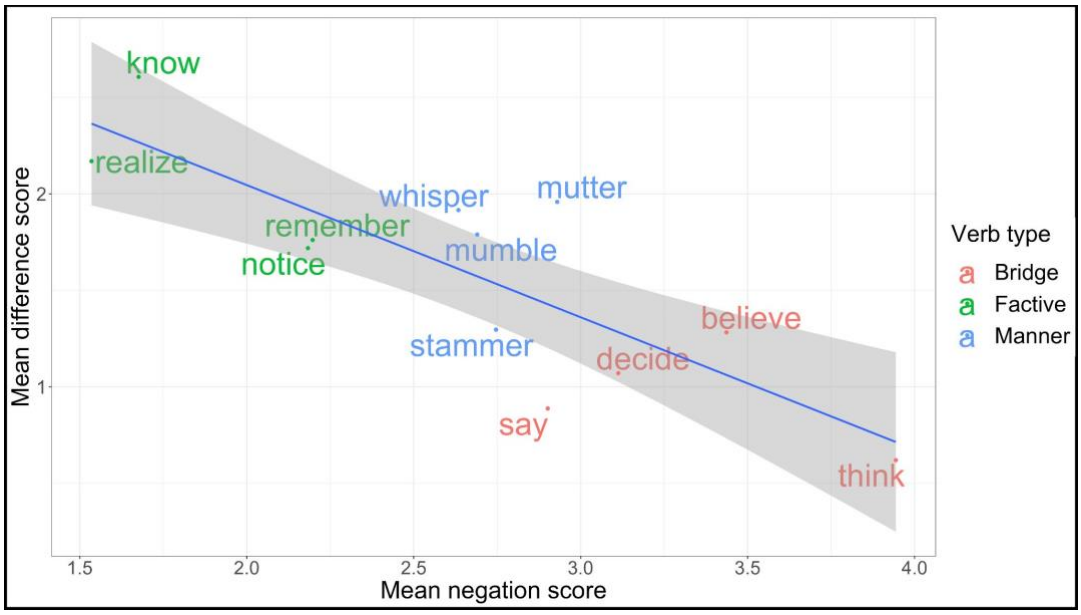
Results of the negation-acceptability analysis of A&G (2008)

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.

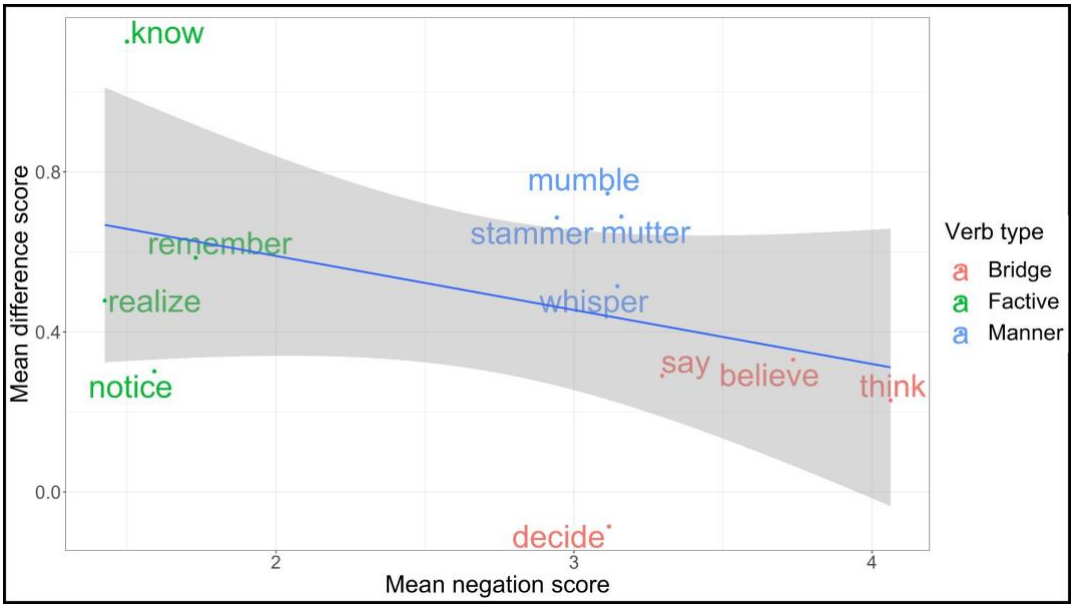
The lower correlations that we observed appear to be derived from at least two sources: first, manner-of-speaking verbs have highly variable difference scores, but very similar negation scores; and second, factive and manner-of-speaking verbs have overall similar difference scores but very different 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).

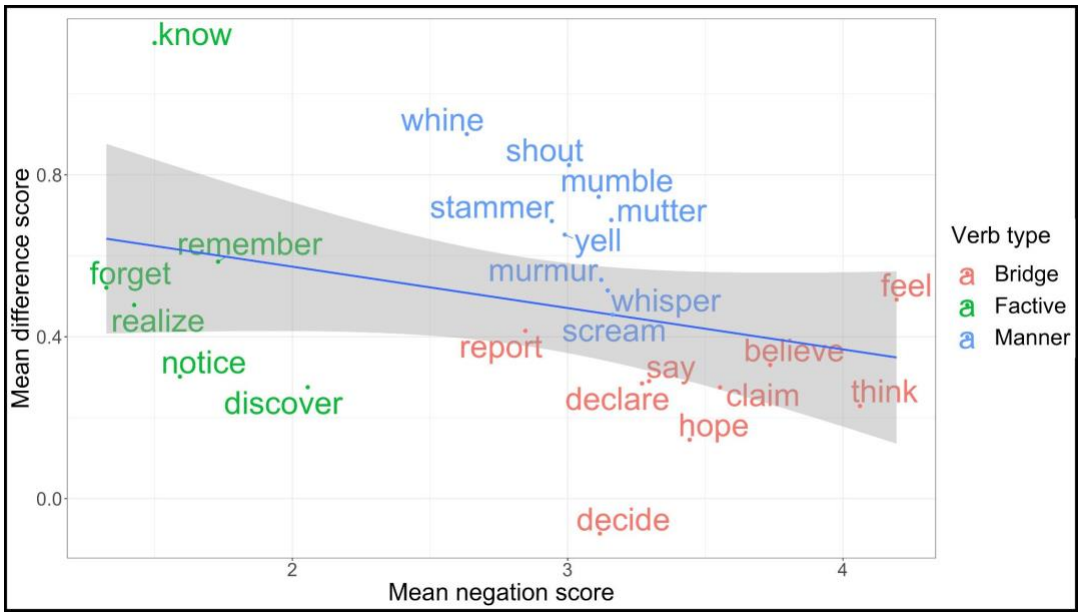


Figure 3. Correlation between mean difference scores and mean negation test scores by verb in A&G (2008) and in the present study (Experiment 1).

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 1: (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.

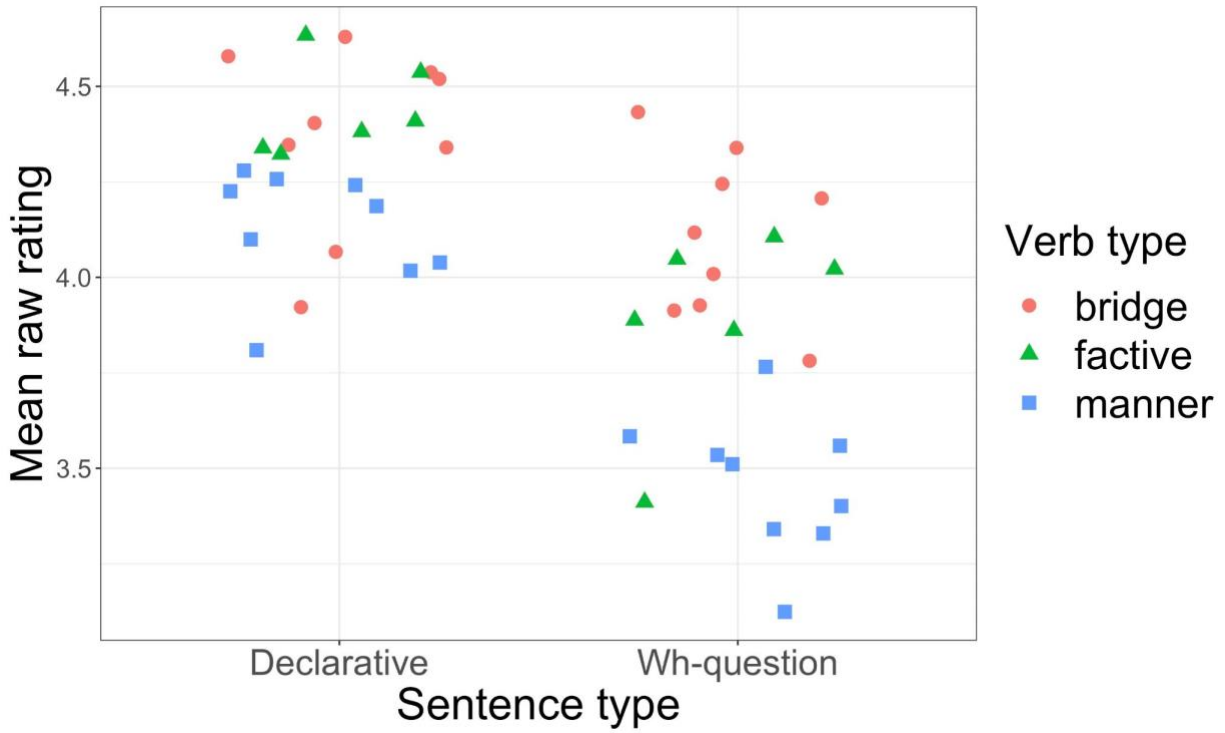


Figure 4. Mean ratings of wh-questions and declarative clauses by verb in Experiment 1, jittered for visualization purposes, for comparison with predictions of the syntactic and semantic accounts in Figures 2a and 2b.

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.

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]

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).

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

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.

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].

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

[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.

[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.

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.

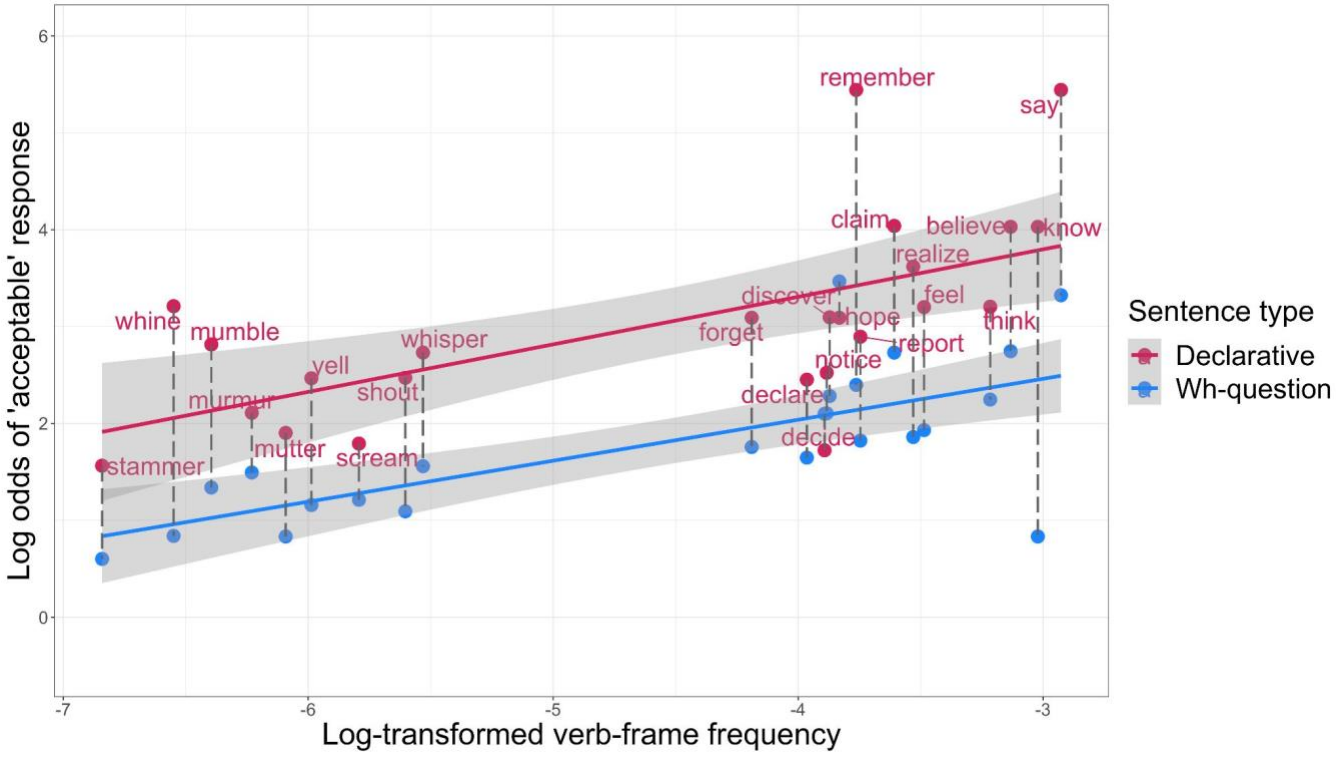


Figure 6. Results of Experiment 1: converted log odds of 'acceptable' response for wh-questions and declarative clauses (transformation of rating 1-2 to 0 and rating 3-5 to 1) against log-transformed verb-frame frequencies by verb. The dashed lines link the two instances of each verb.

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.

2.4 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

[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.

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.

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]

3.1 Participants

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

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,

[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).

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?

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.

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

[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?').

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 ($\beta=0.59$, $z=3.95$, $p<0.001$). There was also a significant main effect of sentence type: declaratives were rated more acceptable than wh-questions ($\beta=-2.45$, $z=-7.88$, $p<0.001$). 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.

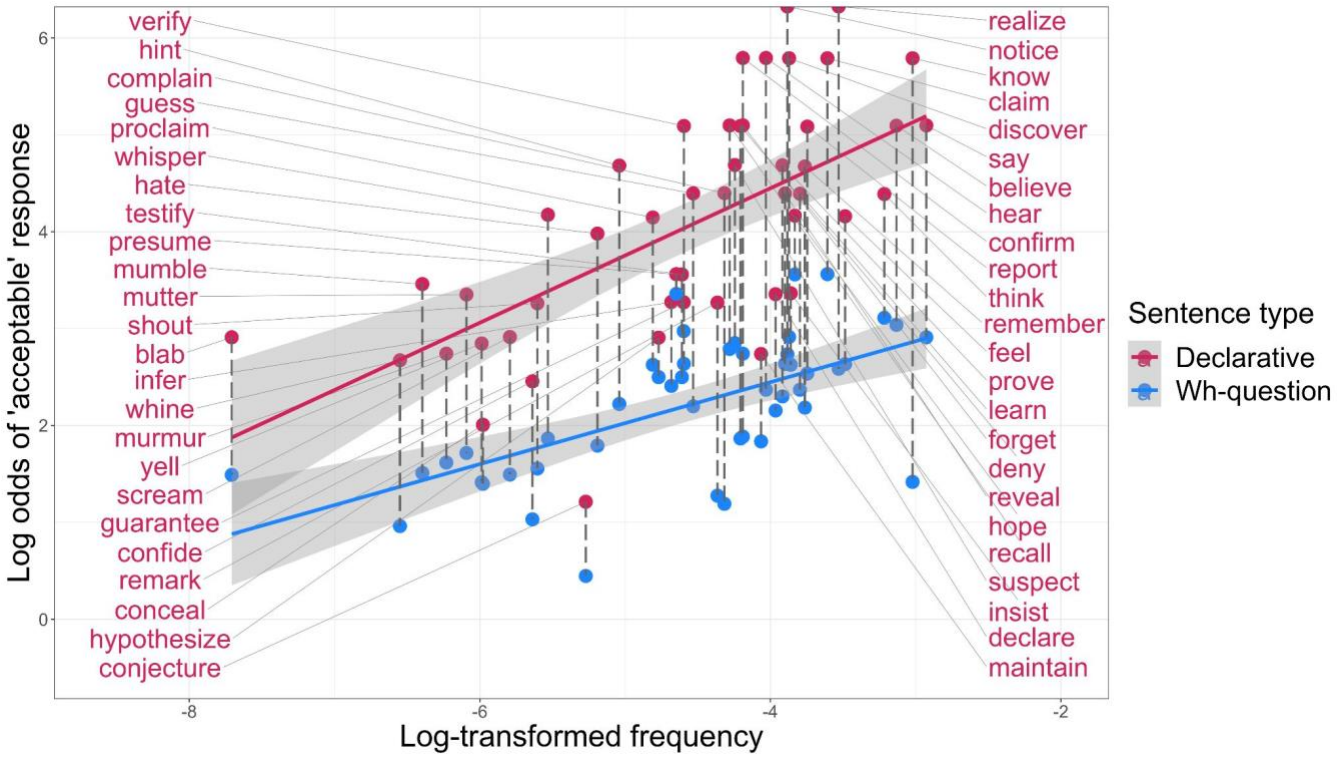


Figure 7. Results of Experiment 2: log-odds of 'acceptable' response for wh-questions and declarative clauses against log-transformed frequencies by verb (48 verbs). The dashed lines link the two instances of each verb.

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 ($r = 0.92$, $t = 15.9$, $p<0.0001$).

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 and Chaves (2020) present an analysis in response to Liu et al. (2019) (a precursor to the current paper). They showed that, once the interaction between verb subcategorization bias and verb type is entered into the model (in addition to these two main effects), the effect of verb subcategorization bias on sentence acceptability 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.

But critically, Richter & Chaves analyze verb subcategorization bias, or the probability of a sentence complement given a verb, as opposed to verb-frame frequency which corresponds to the joint probability of a verb and a sentence complement and is the measure presented in Liu et al. (2019) and here (see Section 2.6). In order to evaluate Richter & Chaves's hypothesis, we obtained their verb frequency and subcategorization measures from the OSF website linked in their paper. First, we compared their measures to our verb-frame frequency measures, which were obtained from a different corpus. In order to do this, we multiplied Richter & Chaves's s-complement subcategorization measure for each verb (e.g., .131 for "say"; .440 for "think"; .328 for "presume", etc.) by an estimate of each verb's frequency (e.g., 200848 for "say"; 59,381 for "think"; 253 for "presume"). We then divided this number by an estimate of the relative frequency of the most frequent verb in Richter & Chaves's verb set ("say", whose frequency of occurrence is approximately .002 in all recent years of Google books), and took a log of the resulting probability in order to get a number that is proportional to our log verb-frame frequency measure. The correlation between this measure and our verb-frame frequency measure (estimated from Google books) for the 45 verbs from Experiment 2 that were in Richter & Chaves' verb set was .923, suggesting that we are measuring similar things in our corpora as Richter & Chaves did.

Using the acceptability data from Experiment 2, we conducted analyses similar to Richter & Chaves using the following glmer formula: glmer(acceptability ~ subcat_bias*sentence_type + (1 + subcat_bias*sentence_type | participant) + (1 + sentence_type | matrix_verb). Subcategorization bias had a significant effect on acceptability (b=3.15, SE = 1.06, p<0.005). The interaction between subcategorization bias and sentence type was not significant (b=0.75, SE = 1.76, p=0.67).
In order to evaluate the relative contributions of subcategorization bias vs. verb frame frequency measure, we then entered both into a logistic regression predicting acceptability in our results from Experiment 2, together with potential interactions with sentence type (wh-question, declarative), using the following glmer formula: acceptability ~ subcat_bias*sentence_type + log_verb_frame_freq*sentence_type+(1 + subcat_bias*sentence_type +log_verb_frame_freq | participant) + (1 + sentence_type | matrix_verb))Verb-frame frequency had a significant effect on acceptability (b=0.51, SE = 0.16, p<0.005), replicating our primary analyses. Subcategorization bias did not have a significant effect (b=1.45, SE = 1.05, p=0.17) and there were no significant interactions with sentence type (subcat:sentence_type: b=2.12, SE = 1.77, p=0.23; verb_frame_freq:sentence_type: b=-0.38, SE = 0.21, p=0.07 .)

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

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)).

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.

4.2 Design and Materials

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

This suggests that verb-frame frequency is a better predictor of acceptability than subcategorization bias. Further details of these analyses are available on OSF.

In addition to using a different notion of frequency than we did in order to attempt to explain the acceptability of island structures, there are some other issues with Richter & Chaves' analyses. 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. And second, 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 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), 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.

sentence with a binary rating scale. Each sentence was followed by a comprehension question (e.g., 'Does this sentence mention an apple?')[15].

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 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.

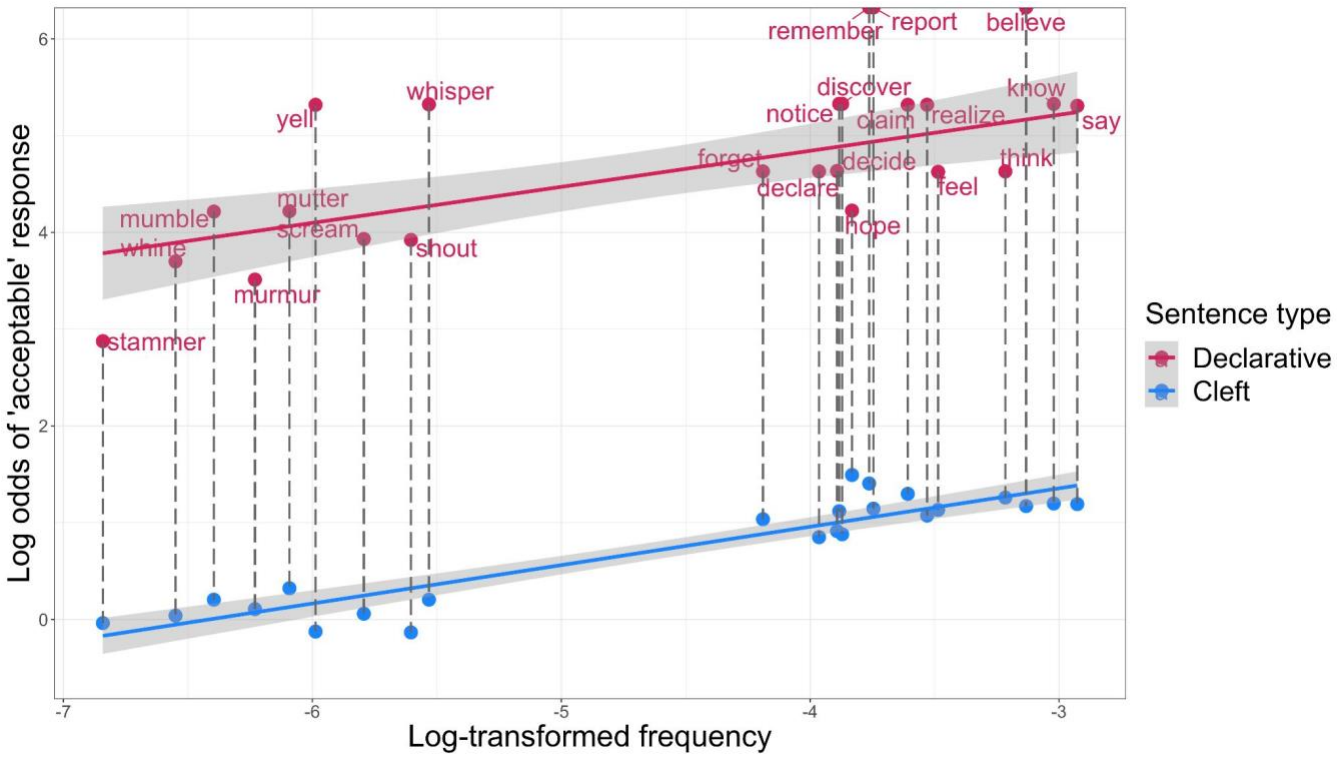


Figure 8. Results of Experiment 3: Log-odds of 'acceptable' response for clefts and declaratives against log-transformed frequencies (24 verbs). The dashed lines link the two instances of each verb.

[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).

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]

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]

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.

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

[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).

[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.

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).

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.

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.

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).

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).

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

Acknowledgement:
The work was supported by a grant from the National Science Foundation Linguistics Program (Award 1534318) to E. Gibson.

Appendix A:

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:

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

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*.

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.

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

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

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

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

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

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.z

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

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.

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
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.
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.
Hoekstra, T., & Kooij, J. G. (1988). The innateness hypothesis. In John A. Hawkins (ed.), *Explaining language universals*, 31-55. Oxford, UK: Blackwell.
Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. MIT Press.
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.
Kothari, A. (2008). Frequency-based expectations and context influence bridge quality. In In M. Grosvald, & D. Soares (Eds.), *Proceedings of WECOL 2008*. UC Davis Department of Linguistics; 2008. http://www.stanford.edu/~anubha/publications.html.
Kush, D., Lohndal, T. and Sprouse, J. (2019). On the Island Sensitivity of Topicalization in Norwegian: An experimental investigation. *Language*, *95*(3), 393-420.
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
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., 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.
MacWhinney, B. (1977). Starting Points. *Language*, *53*(1), 152-168. https://doi.org/10.2307/413059
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.
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.
Phillips, Colin. (2006). The real-time status of island phenomena. *Language*, *82*(4), 795–823. https://doi.org/10.1353/lan.2006.0217
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
Sag, I. A. (2010). English filler-gap constructions. *Language*, *86*(3), 486–545. https://doi.org/10.1353/lan.2010.0002
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
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.
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 of language 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
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

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