Investigating the Role of Verb Frequency in Factive and Manner-of-speaking Islands

Stephanie N. Richter (snrichte@buffalo.edu)
Department of Linguistics, University at Buffalo
Buffalo, NY 14260 USA

Rui P. Chaves (rchaves@buffalo.edu)
Department of Linguistics, University at Buffalo
Buffalo, NY 14260 USA

Abstract

Frequency plays a central role in human cognition, and in language processing in particular. There is growing evidence that acceptability judgements are shaped by the statistics of the input. In this paper, we focus on a type of constraint operative in long-distance dependencies (e.g. *wh*-questions, relative clauses, topicalizations, etc.) which has been claimed to result from verb subcategorization frequency effects. We take a closer look at this hypothesis, and conclude that it does not account for the sentence acceptability contrasts. Rather, the evidence we find suggests that the acceptability of these dependencies hinges on clause-level semantic-pragmatic factors.

Keywords: Frequency effects; Sentence processing; Sentence acceptability; Long-distance dependencies; Neural modelling

Introduction

One property of human language is that words which go together in meaning can occur far away from each other, across many clausal boundaries, as illustrated in (1). In these examples, the *wh*-expression between brackets is somehow interpreted as if it had been 'extracted' from a position immediately after the verb *hates*. We signal the *in situ* position of the extracted phrase via '_', and refer to it as the 'gap' site.

(1) a. [Who] I think Kim said you hated _ was that guy.

b. [Who] do you think Kim said that you hated _?

Although such long-distance dependencies between the *wh*-phrase and the gap can usually cross multiple clausal boundaries, there are certain syntactic configurations which hamper extraction, known as 'islands' (Ross, 1967). In particular, Ross noted that factive verbs (e.g. *know*, which presuppose the truth of their sentential complements) and manner-of-speaking verbs (e.g. *whisper*, which describe physical ways of speaking) hamper long-distance extraction when compared with other verbs, as in the contrast in (2a,b). The verbs in (2b) permit extraction and therefore are referred to as bridge verbs.

(2) a.*[Where/When] did you **know/whisper** [that Kim first kissed Sam _]?
(cf. *Did you know/whisper* [*that Kim first kissed Sam at the movies / last year*]*?*)

b. [Where/When] did you **think/suppose** [that Kim first kissed Sam _]?
(cf. *Did you think/suppose* [*that Kim first kissed Sam at the movies / last year*]*?*)

The non-extracted counterparts given between parentheses are perfectly well-formed, which indicates that the oddness of (2a) is due to the extraction of the *wh*-word over the factive/manner-of-speaking verb. Interestingly, extracting adverbial *wh*-phrases (e.g. *where* and *when*) over factive and manner-of-speaking verbs, as in (2), causes a stronger decline in acceptability than extracting argument *wh*-phrases (e.g. *who* and *what*) over factive and manner-of-speaking verbs, as seen in (3). However, since most studies have focused on the latter case, we shall follow suit, for ease of comparison.

(3) a. ??[What] did Kim **know/whisper** that Bob saw _?
(cf. *Kim knew/whispered that Bob saw something*)

b. [What] did Kim **think/suppose** that Bob saw _?
(cf. *Kim thought/supposed that Bob saw something*)

Whether such extraction 'island' constraints are grammar-based (due to syntax, semantics, and/or pragmatics) or processing-based (due to violated expectations, working memory limitations, and/or contextualization difficulty) remains controversial (Sprouse & Hornstein, 2013). Liu, Ryskin, Futrell, and Gibson (2019) propose that the acceptability of such long-distance extractions over factive, manner-of-speaking, and bridge verbs is best captured in terms of the frequency of verb phrase structure they appear in, rather than in terms of the semantic and pragmatic differences between these verb classes. In what follows, we discuss strengths and weaknesses of previous work, and present new corpus and experimental evidence.

Previous work

Kothari (2008) and Liu et al. (2019) show that the acceptability of factive and manner-of-speaking island phenomena is graded, and that there is an overlap in acceptability between long-distance dependencies involving factive and bridge verbs. These findings contradict syntactic accounts, which predict non-overlapping acceptability. Instead, evidence is more consistent with an account in which pragmatics plays a key role (Kroch, 1998; Erteschik-Shir, 2006; Oshima, 2007; Abrusán, 2011; Abrusán, 2014); e.g., Erteschik-Shir (2006) notes that if it is known that a speaker has a lisp, then (4) is more acceptable because the relevant action is given further context by the embedded clause, not the matrix verb.

(4) [What] did Mike Tyson **lisp** he'd do _?

Table 1: The 136 matrix verbs in the sample

Type	Verbs
Bridge	believe, claim, comment, decide, declare, establish, feel, hope, remark, reply, report, respond, say, think, write
Factive	conceal, discover, forget, hate, ignore, know, learn, note, notice, realize, recall, recollect, regret, remember, resent
Manner	growl, holler, hoot, moan, mumble, murmur, mutter, scream, shout, shriek, stammer, wail, whine, whisper, yell
Other	accept, acknowledge, add, affirm, agree, allege, announce, answer, anticipate, argue, assert, attest, bet, boast, brag, calculate, caution, certify, complain, concede, confess, confirm, consider, deduce, demonstrate, deny, determine, disclose, doubt, dream, emphasize, estimate, expect, explain, fear, find, gloat, guarantee, guess, hear, hint, hypothesize, imagine, imply, indicate, infer, insist, intimate, joke, like, maintain, mention, muse, observe, opine, perceive, plead, predict, presume, pretend, proclaim, promise, propose, prove, reason, reckon, recognize, reiterate, repeat, request, reveal, see, sense, show, signal, signify, speculate, state, suggest, suppose, suspect, swear, testify, theorize, trust, understand, verify, vow, warn, wonder, worry

A view favoring pragmatics is supported by sentence acceptability and self-paced reading evidence that contextualization and semantic priming can weaken these island effects (Kothari, 2008). Kroch (1998), Oshima (2007) and Abrusán (2011) argue that extracting a *wh*-phrase from a factive sentential complement is odd because of a conflict between the semantics of interrogatives and the factivity of the verb, which renders such sentences pragmatically infelicitous. Conversely, for Erteschik-Shir (2006) and Goldberg (2013), manner-of-speaking island effects (among several other islands) are due to a general constraint which hampers extraction from clauses that are not at-issue (i.e., pragmatically backgrounded). Indeed, Ambridge and Goldberg (2008) found a strong correlation ($r = -0.83$, $p < 0.001$) between the acceptability of factive and manner-of-speaking island violations and the degree to which the matrix verb can be construed to express the main action of the clause or not, according to the classic *lie test* of Erteschik-Shir and Lappin (1979).[1] More recently, Tonhauser, Beaver, and Degen (2018) found evidence that whether speakers commit to the content expressed by subordinate clauses is a matter of degree, as it depends on a number of factors, including the prior probability of the event that is described.

There is an alternative account of factive and manner-of-speaking islands that has to do with the frequency with which such verbs occur in a sentential complement frame. Kothari (2008) showed that the higher a verb's bias for the sentential complement frame (computed by dividing the verb's sentential complement frequency by its lemma frequency in corpora), the more acceptable the long-distance dependency, suggesting that the acceptability contrasts found in such dependencies might simply reflect language users' expectations of the typical usage patterns of a verb; see Dąbrowska (2008) for a related account. As Dąbrowska (2004) notes, manner-of-speaking verbs are predominantly used in subcategorization frames which do not have a sentential complement. Along similar lines, Liu et al. (2019) argues that factive and manner-of-speaking island effects are caused by the combination of two factors: the presence of a long-distance dependency, and the frequency of the sentential complement verb frame for the given verb. In other words, the reason why constructions containing bridge verbs are more acceptable than others is because they appear with more frequency than other verbs. In our view, there are several potential problems with Liu et al. (2019). First, the frequency of the sentential complement verb frame was estimated by counting the number of times each verb in their 48-verb study appeared in the Google Books corpus when followed by the complementizer *that*. This is not optimal because the complementizer is optional and, moreover, not all verbs are equally biased for *that*-sentential complements and simple sentential complements. Second, each participant saw 288 sentences in each list plus comprehension questions, and so fatigue is a concern. Third, participants rated each sentence using a rather coarse binary scale ('acceptable' vs. 'unacceptable'). Fourth, no distractor sentences were used, as far as we are aware.

Our goal in this work is to examine the effect of frequency in a larger sample of verbs than those considered by past experiments, using a more accurate estimate of the sentential complement verb frame bias, as well as employing a more fine-grained sentence acceptability design (a 7-point Likert scale instead of a binary scale, or a 5-point scale), while avoiding fatigue and confounds due to a lack of distractors.

Corpus Study

We constructed a list of 136 matrix verbs that have a sentential complement subcategorization frame, shown in Table 1. This list contains 15 bridge verbs, 15 factive verbs, and 15 manner-of-speaking verbs, as retrieved from Liu et al. (2019), previous literature, and our own judgements. We shall refer to these as Bridge, Factive, and Manner verbs, respectively. The remaining 91 verbs are not clearly Factive or Manner verbs, and so we refer to these as Other verbs.[2]

Next, we selected two separate sources for estimating the relative frequency of the sentential complement verb frame for each of our 136 verbs. The first source is the VALEX

[1] Liu et al. (2019) did not replicate the correlation, as acceptability ratings clustered between 4/5 and 5/5 for 74.06% of the items.

[2] Excluded from our sample of 136 verbs are verbs such as *admit* which have radically different meanings when taking direct NP complements (e.g. *We admitted this student* vs. *We admitted that we lied*). It is not clear if previous work controlled the meaning in this manner. Also excluded were verbs like *convince*, which take both an NP complement and a sentential complement (e.g., *I convinced the bouncer that we were in a band*).

Table 2: Corpus study SCR results

Verb type	COCA	VALEX
Factive	0.148 (*SD* = 0.153)	0.179 (*SD* = 0.121)
Manner	0.016 (*SD* = 0.009)	0.039 (*SD* = 0.012)
Bridge	0.335 (*SD* = 0.16)	0.260 (*SD* = 0.134)

database (Korhonen, Krymolowski, & Briscoe, 2006), a large subcategorization lexicon for English which includes subcategorization frequency information for 6,397 verbs, based on approximately 16 million sentences. We extracted frequency data from frames which took a sentential complement in finite or base form from sublexicon #4, selected for its lack of smoothing. The second source consisted of a random sample of 2.5 million sentences extracted from the Corpus of Contemporary American English (COCA) (Davies, 2008–), tagged and parsed with Stanford CoreNLP (Manning et al., 2014) ver. 3.9. We used Tregex (Levy & Andrew, 2006) to search for the same verb usages as in the VALEX frames.

Following Kothari (2008), we obtained two counts for each verb lemma: the frequency of the verb (however inflected) when followed by a sentential complement with or without the complementizer *that* (direct quotations were excluded), and the frequency of the verb (however inflected), no matter the construction. Using these frequencies, we compute each verb's bias for the sentential complement frame; we will refer to this as the *Sentential Complement Ratio* (SCR), computed as such in (5), where SC = sentential complement.

$$\text{(5)}\quad \text{SCR}_{\text{lemma}} = \frac{\#(\text{verb lemma used with SC})}{\#(\text{verb lemma})}$$

Liu et al. (2019) argue that *wh*-interrogative constructions with bridge verbs are more acceptable than others because bridge verbs are more frequent. If this is correct, then Factive and Manner verbs should have the lowest SCR values.

Results

The results are shown in Table 2 and illustrated in Figure 1. As expected, the SCRs of Manner and Factive verbs are lower than those of Bridge verbs. However, there is an SCR overlap between Bridge verbs and Factive verbs, and thus there are various Bridge verbs with lower SCRs than those of Factive verbs. With regard to the Other category, we identified 13 verbs which have lower SCRs than most Factive verbs in the sample, and which may be best characterized as bridge verbs, as illustrated by the acceptability of (6). A list of such verbs is shown in Table 3. The raw frequency of each verb lemma used within the sentential complement frame is $Freq_{SC}$, and the raw frequency of the lemma is $Freq_V$.

(6) a. [What] did you **establish** [that Don stole _]?

b. [When] did you **establish** [that Don stole the car _]?

In particular, 7 of the Bridge verbs with the lowest SCRs overlap with the SCR range of Manner verbs (mean 0.038, *SD* = 0.01). A similar pattern arises for the VALEX SCRs of verbs like *repeat*, *expect*, *perceive* and *add*, which are extremely low as well. Bridge verbs with such low SCRs should not exist, according to the proposal of Liu et al. (2019).

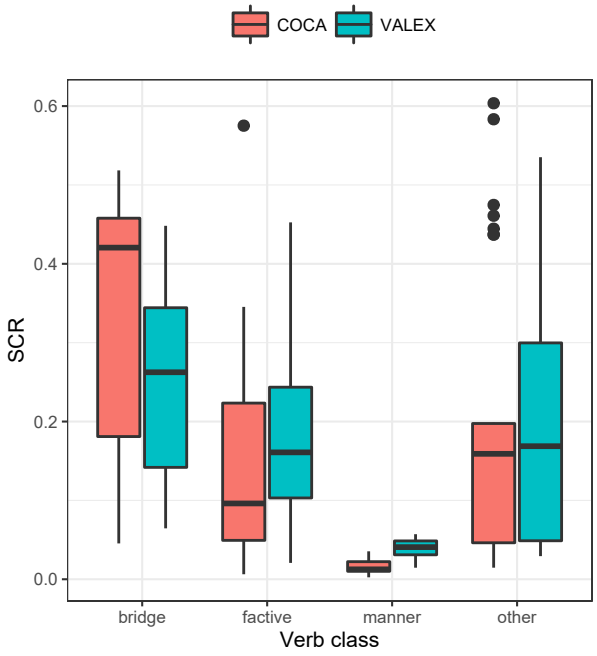

Figure 1: SCRs across verb classes, in both datasets

Experiment

We selected all 15 Factive and 15 Manner verbs from our sample of 136 verbs, and three other groups of verbs from the sample, containing 15 verbs each, characterizing different ranges of SCR values. These remaining verbs are thus a combination of Bridge and Other verbs. We thus have 15 verbs in the low-range of SCR, 15 verbs in the high-range, and 15 verbs in the near-average SCR range. We refer to these verbs

Table 3: Verbs with low SCR (COCA)

Verb	Freq_{SC}	Freq_V	SCR
like	1135	20136	0.056
add	707	13133	0.053
consider	491	9129	0.053
perceive	39	750	0.052
expect	490	9844	0.049
comment	52	1144	0.045
reply	111	2667	0.041
wonder	228	7053	0.032
establish	93	2907	0.031
write	480	18385	0.026
repeat	47	2220	0.021
answer	98	5021	0.019
respond	60	4085	0.014

as Low_{SCR}, High_{SCR}, and Mid_{SCR}, respectively. In the experiment below, we compare the SCRs of these 75 verbs in our sample with human sentence acceptability judgements.

We also conducted a norming experiment with the declarative counterparts of our interrogative items, in order to control for the effect on sentence acceptability that the choice of verb may have independently of the presence of a long-distance dependency. Our primary experiment with *wh*-interrogatives contained sentences such as (7a), while the norming experiment consisted of their counterparts as in (7b).

(7) a. What did Tom know/whisper/say that Henry saw?

b. Tom knew/whispered/said that Henry saw something.

According to Kothari (2008) and Liu et al. (2019), the sentential complement verb frame bias should be highly correlated with the acceptability difference between (7a,b).

Method

Participants We analyzed data from 161 participants with IP addresses originating from the United States. The participants were recruited through Amazon.com's Mechanical Turk (AMT) crowdsourcing marketplace, and their responses were collected using Ibex Farm (Drummond, 2013). The participants had comprehension question accuracy levels above 75% and self-reported as native speakers of English at the end of the experiment. Participants were informed that self-reporting as non-native would not affect their compensation.

Design and materials There were 5 conditions (Factive, Manner, High_{SCR}, Mid_{SCR}, Low_{SCR}), based on the corpus study discussed above, and 15 verbs per condition. There were thus 5 versions of each experimental item, as the example in (8) illustrates, pseudo-counterbalanced across 5 lists so that each participant only responded to one version of each experimental item. Although there were 5 lists, each list contained a total of 6 verbs from each of the 5 conditions, in equal proportions, so that each verb appeared in two lists.

(8) What did Ann $\begin{pmatrix} \text{discover}_{(\text{Factive})} \\ \text{holler}_{(\text{Manner})} \\ \text{believe}_{(\text{High})} \\ \text{decide}_{(\text{Mid})} \\ \text{answer}_{(\text{Low})} \end{pmatrix}$ that Joe bought?

This design was adopted to avoid having only 3 items per condition across the five lists, which would have been arguably too low, while at the same time avoiding an excessively long experiment, which would be prone to causing fatigue.

The items were pseudo-randomized with 45 distractors, for a total of 75 sentences per list. There were three types of distractor, as illustrated in (9). One third of the distractors were moderately acceptable and were potentially followed by a comprehension question, as in (9a); one third had low acceptability like (9b), and the remaining were odd, as in (9c).

(9) a. What made the sound that Jake heard? (Good)
[T/F Query]: Jake is most likely not deaf. [T]

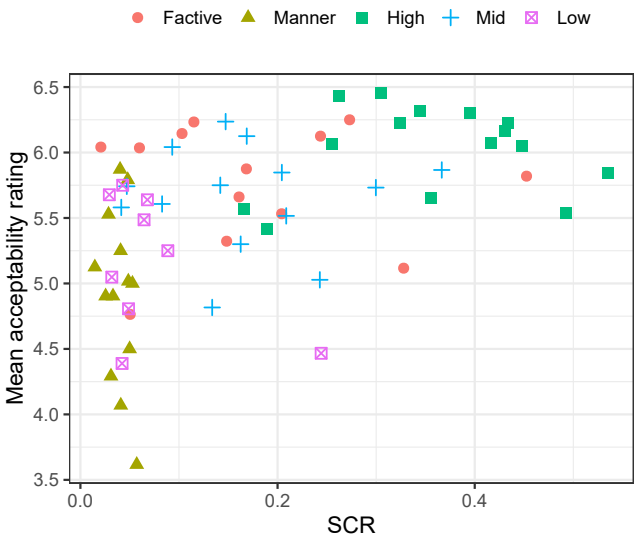

Figure 2: Mean acceptability ratings of *wh*-interrogatives vs. their corresponding matrix verbs' SCR values in VALEX

b. What does Obraya ask that Elizot do? (Passable)

c. What did Martin mean that Teresa married? (Odd)

The norming experiment was conducted with the declarative counterparts of the items, as illustrated by (7). The declarative items were counterbalanced across 5 lists in exactly the same way as in the interrogative experiment counterpart, interspersed with 45 distractor sentences and pseudo-randomized so that no two participants saw the same sentence order. Data from a different group of 155 self-reported native speakers recruited via AMT was collected. These participants were tasked to rate how natural the declarative sentences were, using a 7-point Likert scale, and to answer the comprehension questions with at least 75% accuracy.

The acceptability ratings were generally quite high, clustering within a very narrow interval between 5.5 and 6.5. More specifically, the declarative Factive items received a mean acceptability rating of 6.03 (SD = 1.21); Manner, 5.82 (SD = 1.27); High_{SCR}, 6.08 (SD = 1.12); Mid_{SCR}, 5.88 (SD = 1.21); Low_{SCR}, 5.6 (SD = 1.46). We will include these ratings as a predictor in our analysis of the effect of SCR on the acceptability of the interrogative counterparts. Additionally, 'Good' distractors had a mean of 5.7 (SD = 1.6), 'Passable' received 4.2 (SD = 2.05), and 'Odd' received 2.56 (SD = 1.6).

Procedure Participants were asked to judge how natural each *wh*-interrogative was by giving each a rating from 1 ('very unnatural') to 7 ('very natural'). To ensure that comprehenders were attending to the structure and meaning of the experimental items, half of the 'Good' distractors were followed by a True/False question as illustrated in (9a), on a different screen. Participants were informed when their answers were incorrect, and the overall accuracy ratio was 85%.

Results

The mean response for Factive items was 5.75 ($SD = 1.3$); for the Manner, 4.96 ($SD = 1.7$); High$_{SCR}$, 6.01 ($SD = 1.15$); Mid$_{SCR}$, 5.64 ($SD = 1.31$); Low$_{SCR}$, 5.1 ($SD = 1.6$). 'Good' distractors had a mean response of 5.71 ($SD = 1.6$), 'Passable' had a mean of 4.2 ($SD = 2.05$), and 'Odd' had a mean of 2.56 ($SD = 1.6$). The results for the VALEX dataset are depicted in Figure 2, and those for the COCA dataset are very similar as shown in Figure 3.

Ordinal Mixed-Effect Regression (OMER) models were fit with the VALEX SCR and the COCA SCR as independent variables, using the `ordinal` package (Christensen, 2018). The intercept was allowed to be adjusted by the (centered) declarative norming ratings, subjects, items, and lists, in order to account for random effects. Both models were significant: $\beta = 4.2$ ($SE = 0.72$, $z = 5.78$, $p < 0.0001$) for VALEX, and $\beta = 3.6$ ($SE = 0.46$, $z = 7.68$, $p < 0.0001$) for COCA. From this result, one could conclude that SCR is a good predictor of extraction acceptability. However, it is clear from Figure 2 that verbs of different classes cluster very differently. For example, Manner verbs vary widely in their acceptability, yet are clustered within low SCR values. Factive verbs, meanwhile, show the opposite: they mostly all have relatively high acceptability ratings, yet vary wildly in SCR value; see the standard deviations (SD) reported in Table 2. Verbs categorized as High and Mid also appear to vary in SCR to a large degree, while Low verbs do not show this same tendency. In the COCA dataset, the vertical SCR bands are even more pronounced for most verb classes and generally span a wide range of acceptability ratings. This is illustrated in Figure 3, which shows the 75 verbs from COCA ranked in decreasing mean acceptability rating. To probe for an effect between each of the verb classes, we next fit OMER models like those above but instead with the interaction between SCR and verb type as the predictor. These models were not significant for VALEX (all p's > 0.25), nor for COCA (all p's > 0.09). Analogous LMER models with the difference between interrogative and declarative acceptability z-scores as the dependent variable were likewise non-significant. In our view, the effect reported above disappears within each class because different verbs are distributed very differently, a problematic result for verb frame frequency-based accounts.

Finally, we probed for a correlation between the frequency with which our verbs appear in interrogative long-distance dependencies in our COCA dataset and the mean acceptability ratings for items containing the same verb in our sentence acceptability experiment. No significant correlation was found for VALEX or COCA (all p's > 0.6) either.

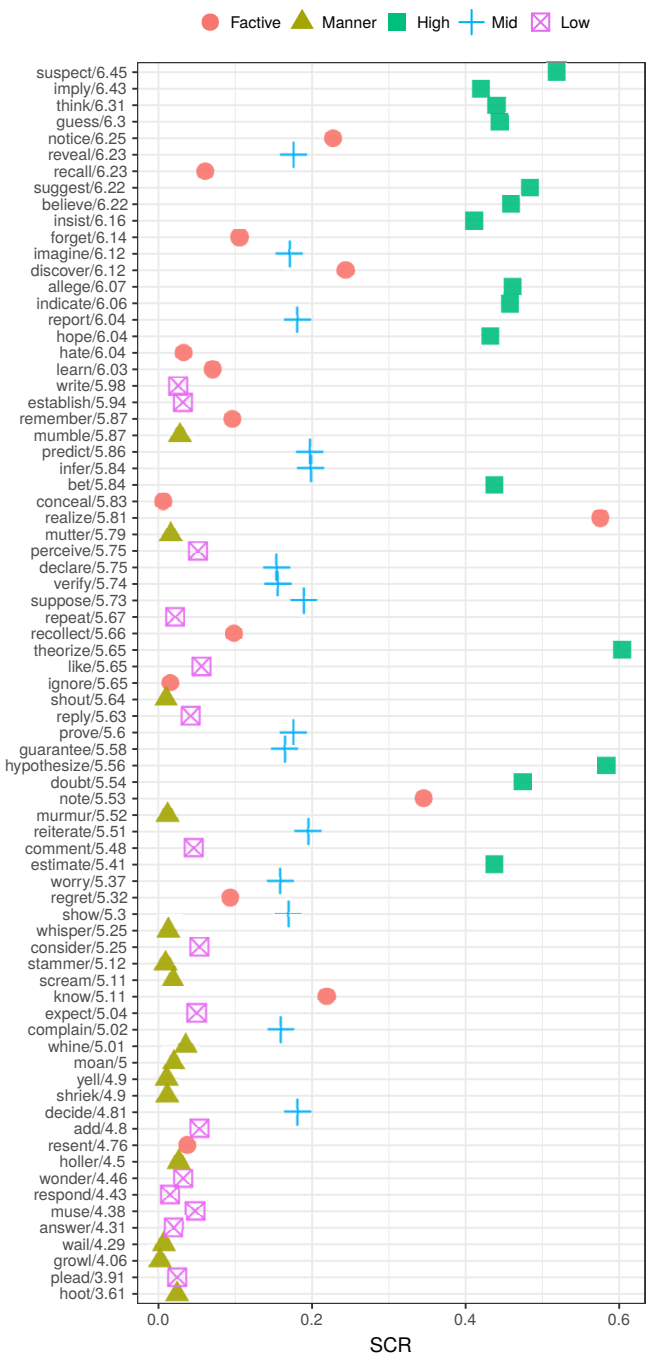

Figure 3: Mean acceptability ratings of *wh*-interrogatives vs. their corresponding matrix verbs' SCR values in COCA

Discussion

In our two datasets, we found bridge verbs that have lower SCRs than island-inducing verbs. We observed an overall effect of SCR on interrogative long distance dependencies acceptability (controlled by the normed declarative rating counterparts), but upon closer inspection it becomes clear that there are many verb classes with similar sentential complement ratios which exhibit a wide range of acceptability ratings, beyond what would be expected if SCR determined acceptability. For example, High$_{SCR}$ verbs should generally yield more acceptable long-distance dependencies than Mid$_{SCR}$ verbs, and the latter should in turn be more acceptable than long-distance dependencies with Low$_{SCR}$ verbs, but what we found was an extensive amount of acceptability rat-

ing overlap across the three verb classes. Overall, these results suggest that the frequency of the sentential verb frame is not the major factor in the acceptability of interrogative extractions from sentential complements, contra Dąbrowska (2008), Kothari (2008), and Liu et al. (2019).

In our view, these results favor of a multivariate account where semantics and pragmatics modulate the acceptability of such constructions in different ways per verb class (Erteschik-Shir, 2006; Oshima, 2007; Abrusán, 2014). For example, it is clear within the VALEX dataset that Factive and Manner islands are not equally strong, and thus any account needs to take into consideration possible sources of gradience. We conjecture that the probability of the utterance as a whole, the probability of the event that is described (given the context), and the likelihood that the content of the embedded clause is at-issue (Tonhauser et al., 2018), are likely contributors to the observed differences in acceptability ratings as measured in a 7-point Likert scale.

Computational modelling

Perhaps factive and manner-of-speaking island effects depend (in part) on the probability of the utterance as a whole, and in particular, on the probability of the event, and the likelihood that the content of the embedded clause is at-issue. Such a result would be consistent with the view that semantics and pragmatics modulate the acceptability of extraction from factive and manner-of-speaking verbs (Erteschik-Shir, 2006; Oshima, 2007; Abrusán, 2014).

Estimating the syntactic, semantic, and pragmatic probability of an utterance poses obvious practical problems; as a proxy we turn to large-scale probabilistic language models; in particular, to OpenAI's GPT-2 (Radford et al., 2019), a state-of-the-art 345-million parameter pre-trained neural network model which is one of the best performers on a range of language modelling tasks (Warstadt et al., 2020), including the modelling of long-distance dependencies very much like the interrogatives that are presently under scrutiny (Da Costa & Chaves, 2020). Although GPT-2's training objective is simply to predict the next word given a preceding context, it has seemingly 'learned' useful linguistic information.

Using GPT-2 as a proxy to estimate sentence probability, we adopt the methodology of Wilcox, Levy, Morita, and Futrell (2018) in which large-scale pretrained recurrent neural networks were probed for their ability to represent long-distance dependencies using surprisal (Hale, 2001; Levy, 2008; Smith & Levy, 2008). The surprisal $S(w)$ of a word w is estimated as the log of the inverse probability of w according to the network hidden state softmax activation h before consuming w, given all previous words in the sentence:

(10) $S(w) = -log_2\, p(w|h)$

We gave the model the same 150 interrogative experimental items as those given to our human participants in the experiment we have reported, one word at a time. At each step, we used the softmax activation of GPT-2 (i.e., the predicted probability of the upcoming word) to compute the surprisal of the next (unseen) word in the input. We summed the surprisal values of every word in each sentence, as in Wilcox et al. (2018), with the goal of capturing distributed, as well as global (sentence-wide) effects. The mean whole-sentence surprisal per verb was compared with the mean sentence acceptability in our experiment, per verb. The result is shown in Figure 4, and suggests a moderate correlation ($r = -0.588$, $t = -6.21$, $p < 0.0001$). These findings suggest that sentence probability (including that of the expressed proposition) plays an important role in the acceptability of extraction from factive and manner-of-speaking verbs, an outcome consistent with Erteschik-Shir (2006), Goldberg (2013), and Tonhauser et al. (2018). Further work is necessary to test this.

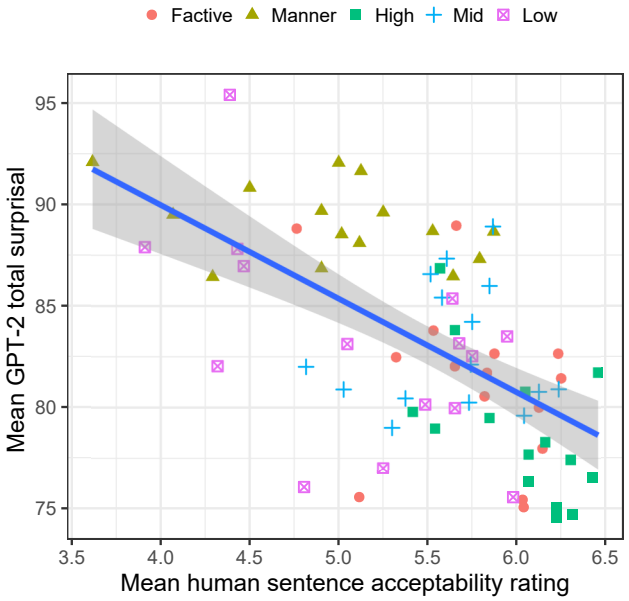

Figure 4: GPT-2's mean whole-sentence surprisal vs. human interrogative sentence ratings, for each verb in the experiment

Conclusion

Our results suggest that verb frame frequency does not account for the acceptability of extraction from sentential complements. In a large corpus study, we found various verbs that are as infrequent as many or most Factive and Manner verbs, but which nonetheless do not hamper long-distance dependencies. Although we found an apparent correlation between sentential complement bias and the sentence acceptability of *wh*-phrase extractions from such complements, we found that this effect disappears within each verbal class. Our findings indicate that different verb classes cluster differently, in ways which are unexpected for the sentential complement bias hypothesis. Finally, we found a moderate correlation between overall sentence probability and human acceptability ratings, a result consistent with accounts where graded semantic/pragmatic factors play a role (Oshima, 2007; Abrusán, 2011; Goldberg, 2013; Tonhauser et al., 2018).

All relevant materials and code used for this paper are available online at `https://osf.io/ndvc4/`.

Authorship contribution statement

S.R. extracted COCA and VALEX data, and constructed Ibex experiment; R.C. did statistical analyses and the GPT-2 experiment. R.C. and S.R. created plots and wrote the paper.

Acknowledgements

We are very thankful to the CogSci reviewers for their feedback, as well as to Jean-Pierre Koenig, Adele Goldberg, and the SynSem group at UB for comments and suggestions.

References

Abrusán, M. (2011). Presuppositional and negative islands: A semantic account. *Natural Language Semantics*, *19*, 257–321.

Abrusán, M. (2014). *Weak island semantics*. Oxford: Oxford University Press.

Ambridge, B., & Goldberg, A. E. (2008). The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, *19*, 357–389.

Christensen, R. H. B. (2018). *Ordinal-regression models for ordinal data.* Retrieved from http://www.cran.r-project.org/package=ordinal/ (R package version 2018.8-25)

Da Costa, J. K., & Chaves, R. P. (2020). Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics (SCiL)* (Vol. 3, pp. 20–30). New Orleans, LA: Society for Computation in Linguistics.

Dąbrowska, E. (2004). *Languge, mind, and brain: Some psychological and neurologial constraints on theories of grammar*. Edinburgh: Edinburgh University Press.

Dąbrowska, E. (2008). Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics*, *19*(3), 391–425.

Davies, M. (2008–). *The Corpus of Contemporary American English (COCA): 600 million words, 1990–present.* (Available online at https://www.english-corpora.org/coca/)

Drummond, A. (2013). *Ibex 0.3.7 manual.* (Ms., [available at: http://spellout.net/latest_ibex_manual.pdf])

Erteschik-Shir, N. (2006). Bridge phenomena. In M. Everaert, H. V. Riemsdijk, R. Goedemans, & B. Hollebrandse (Eds.), *The Blackwell Companion to Syntax Volumes* (Vol. 5, pp. 284–294). Oxford: Blackwell.

Erteschik-Shir, N., & Lappin, S. (1979). Dominance and the functional explanation of island phenomena. *Theoretical Linguistics*, *6*, 41–86.

Goldberg, A. E. (2013). Backgrounded constituents cannot be extracted. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 221–238). Cambridge: Cambridge University Press.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001* (pp. 159–166). Pittsburg, PA: ACL.

Korhonen, A., Krymolowski, Y., & Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation.* Genova, Italy.

Kothari, A. (2008). Frequency-based expectations and context influence bridge quality. In M. Grosvald & D. Soares (Eds.), *Proceedings of WECOL 2008: Western Conference on Linguistics* (pp. 136–149). Fresno, CA: California State University.

Kroch, A. (1998). Amount quantification, referentiality, and long *wh*-movement. In *University of Pennsylvania Working Papers in Linguistics* (Vol. 5(2), p. Article 3). Philadelphia, PA: University of Pennsylvania.

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

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation (LREC).* Genoa, Italy.

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, QB: Cognitive Science Society.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60).

Oshima, D. Y. (2007). On factive islands: pragmatic anomaly vs. pragmatic infelicity. In *Proceedings of the 20th Annual Conference on New Frontiers in Artificial Intelligence* (pp. 147–161). Berlin, Heidelberg: Springer-Verlag.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* (https://github.com/openai/gpt-2)

Ross, J. R. (1967). *Constraints on variables in syntax*. Doctoral dissertation, MIT, Cambridge, Massachusetts.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the Thirtieth Annual conference of the Cognitive Science Society* (pp. 595–600).

Sprouse, J., & Hornstein, N. (2013). *Experimental syntax and island effects*. Cambridge: Cambridge University Press.

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.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics* (Vol. 3).

Wilcox, E., Levy, R. P., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*.