The island status of clausal complements: Evidence in favor of an information structure explanation

BEN AMBRIDGE and ADELE E. GOLDBERG*

Abstract

The present paper provides evidence that suggests that speakers determine which constructions can be combined, at least in part, on the basis of the compatibility of the information structure properties of the constructions involved. The relative "island" status of the following sentence complement constructions are investigated: "bridge" verb complements, manner-of-speaking verb complements and factive verb complements. Questionnaire data is reported that demonstrates a strong correlation between acceptability judgments and a negation test used to operationalize the notion of "backgroundedness". Semantic similarity of the main verbs involved to think *or* say *(the two verbs that are found most frequently in long-distance extraction from complement clauses) did not account for any variance; this finding undermines an account which might predict acceptability by analogy to a fixed formula involving* think *or* say*. While the standard subjacency account also does not predict the results, the findings strongly support the idea that constructions act as islands to wh-extraction to the degree that they are backgrounded in discourse.*

Keywords: *island constraints; constructions; sentence complements; manner of speaking verbs; factive verbs; bridge verbs.*

1. Introduction

Imagine the President was given an incriminating top secret FBI file about a person who worked closely with him. Watching him storm out

* We are grateful to Ewa Dąbrowska, Mirjam Fried, Jon Sprouse, Mike Tomasello and Robert Van Valin for comments on an earlier draft. We are also grateful to Blazej Galkowski and to Ewa Dąbrowska for discussion about the Polish facts. This work was supported by an NSF grant to the second author: NSF 0613227. Correspondence addresses: Ben.Ambridge@liverpool.ac.uk, adele@princeton.edu.

Cognitive Linguistics 19–3 (2008), 349–381 0936–5907/08/0019–0349
DOI 10.1515/COGL.2008.013

of the room, the people gathered may well wonder who the report was about. And yet they could not formulate the question as follows:

(1) *Who did he just read the report that was about _?

As this example illustrates, even when questions appear to be semantically appropriate, there are constraints on what can count as a question. Where do such constraints come from? The question has been at the heart of linguistic theorizing for decades. Many researchers assume that the answer must lie in a system of innate linguistic knowledge that is built on purely formal principles that are specific to language, since it is not difficult to come up with contexts in which ill-formed questions would seem to be semantically appropriate as in the example just given (e.g., Chomsky 1973; Ross 1967; Pinker and Bloom 1990).

In this paper we compare the viability of the following proposals: a) a formal ''subjacency'' account, b) an account that predicts acceptability to be determined by semantic comparison to a high-frequency formula, and c) the hypothesis that discourse properties of the constructions involved determine the relative acceptability of long-distance dependencies.

1.1. *Filler-Gap constructions*

WH-questions typically involve a constituent that appears in a position other than its canonical position. We refer to the displaced constituent as the *filler* (indicated by italics), and the place where the constituent would appear in a simple sentence, the *gap* (''_''). In this way, we can avoid the common terminology that the filler is ''extracted'' from the site of the gap and ''moved'' to the front of the sentence, since we do not assume that there is any actual movement (see e.g., Ambridge et al. 2006; Sag and Fodor 1994; Van Valin 1993; Ambridge et al. 2007; for non-movement accounts of simple and complex question formation). An example of a question filler-gap construction is given in (2):

(2) *Who* did she think he saw _?

Relative clauses and topicalizations are other types of filler-gap constructions as in (3) and (4):

(3) I met the man *who* I think you saw _. (relative clause)
(4) *Whitefish and bagels*, she served _ (topicalization)

Ross (1967) first observed constraints on filler-gap relations. Certain syntactic constructions are ''islands'' to such relations: in particular, they

may not contain the gap.[1] Syntactic islands include complex noun phrases, subjects, adjuncts, complements of manner-of-speaking verbs and complements of factive verbs as illustrated below.

Table 1. *Classic examples of "Island" constraints*

*Who did she see the report that was about _? (cf. She saw the report that was about x)	Complex NPs (both noun complements and relative clauses)
*Who did that she knew bother him_? (cf. That she knew x bothered him)	Subjects
??What did she leave the movie because they were eating _? (cf. She left the movie because they were eating x)	Presupposed adjuncts
??What did she whisper that he saw_? (cf. She whispered that he saw x)	Complements of manner-of-speaking verbs
??What did she realize that he saw_? (cf. She realized that he saw x)	Complements of factive verbs

Judgments in the case of complex NPs and subject islands are more robust, and less dependent on context, than in any of the latter three instances. Exploring these subtle differences in judgments requires us to look in a more detailed way at the discourse functions of each of the constructions involved. We return to this issue of graded judgments below.

1.2. *Subjacency*

How should constraints on filler-gap constructions be accounted for? Since Chomsky (1973), the dominant view has been that constraints on filler-gap constructions arise from a "subjacency" constraint: namely that the gap cannot be separated from the filler by two or more "bounding nodes," where S and NP are defined to be bounding nodes.[2] Subjacency is a parade example of a constraint that has been claimed to be formal and specific to language: part of "universal grammar" (Newmeyer

1. The "island" metaphor was based on the idea that the filler moved from the gap position to the front of the sentence. Islands refer to constituents from which a filler cannot move.
2. NP and S are considered bounding nodes in English. NP and S′ appear to be bounding nodes in Italian (Rizzi 1982) and S, S′ and NP appear to be bounding nodes in Russian (Freidin and Quicoli 1989). That is, Italian speakers can apparently extract out of WH-complements, while Russian speakers can only extract out of main clauses.

1991). The subjacency account predicts that complex NPs, subjects and all adjuncts should be islands.

At the same time, the subjacency account predicts that gaps within clausal complements should be acceptable since only one bounding node (S) intervenes between the filler (*who*) and the gap (_). This prediction in fact holds when the main verb is a semantically light ("bridge") verb of saying or thinking (including *think, say, believe*) (cf. 5):

(5) Who did she think that he saw _?

However, while gaps within the complement clauses of bridge verbs are, as predicted, acceptable, the subjacency account does not explain why gaps within the complements of manner of speaking verbs or factive verbs should be less than fully acceptable, since the syntactic structures appear to be the same (Erteschik-Shir and Lappin 1979; Ross 1967).

(6) ??Who did she mumble that he saw_?
Manner of speaking verb complement
(7) ??Who did she realize that he saw_?
Factive verb complement

The natural solution for a syntactic account is to argue that the syntactic structures are not actually the same. In fact it has been suggested that the complements of manner of speaking verbs are adjuncts, not arguments (Baltin 1982). This idea is supported by the fact that the clausal complement is optional:

(8) She shouted that he left.
(9) She shouted.

Since adjuncts are predicted to be islands on the subjacency account, this move predicts that clausal complements of manner of speaking verbs should be islands. However, clausal complement clauses are restricted to appear with a fairly narrow set of verbs including verbs of saying and thinking; this restrictiveness is a hallmark of arguments, not adjuncts. Moreover, (9) does not convey the same general meaning as (8) insofar as only (8) implies that propositional content was conveyed; the change of basic meaning when omitted is another hallmark of arguments. In addition, direct object arguments can replace clausal complements (e.g., 10), and yet it would be highly unusual to treat a direct object as an adjunct:

(10) She shouted (the remark).

Finally, the possibility of treating the complement clause as an adjunct clearly does not extend to factive verbs, since their clausal complements are not generally optional (cf. 11–12).

(11) She realized that he left.
(12) ??She realized.

Kiparsky and Kiparsky (1971) suggest a different solution to account for the island status of clausal complements of factive verbs. They suggest that factive clausal complements contain a silent *the fact* rendering the clausal complements part of a complex NP (as in 13).

(13) She realized *the fact* that he left.

This analysis predicts that the complement clauses of factive verbs should be as strong islands as overt NP complements, since expressions such as (14) and (15) would be structurally identical:

(14) *Who did she realize the fact that he saw?
(15) ??Who did she realize that he saw?

Intuitively, however, (14) is less acceptable than (15). Moreover, positing a silent *the fact* phrase to account for the ill-formedness of examples like (15) is ad hoc unless a principled reason can be provided for *not* positing a silent NP (e.g., *the idea*) in the case of bridge verbs which readily allow extraction.

(16) *Who did she believe the idea that he saw?
(17) Who did she believe he saw?

To summarize, if, in fact, the syntax is the same and only the lexical semantics differs, subjacency does not predict variation in judgments across different verb classes. The complement clauses must be reanalyzed as either adjuncts or parts of complex NPs (to our knowledge, it has not been proposed that they could be subjects, but that would be the other option), but each of these possibilities raises issues that would need to be addressed for the proposed alternative analyses to be convincing.

1.3. *A possible direct-analogy account*

Other researchers have emphasized that long-distance filler-gap constructions are exceedingly rare in spoken corpora. Dabrowska (2004) and Verhagen (2006) both observe that the only long-distance filler-gap expression to occur with any regularity at all are specific formulas with the verb *think* or *say (WH do you think/say S)?*

Dabrowska notes that of a total of 49 long-distance filler-gap constructions produced by five children in CHILDES corpora, all but two were

instances of these formulas. Dabrowska notes further that 96 percent of adult's long-distance filler-gap constructions in the Manchester corpus also involve the main verb *think* or *say* (2004: 197).

Verhagen (2006) likewise notes that in both English and Dutch corpora, questions out of main verb complements are almost uniformly instances of the formula, *WH do you think S?* or, in the case of Dutch, *WH denk-pron*$_{2nd}$ *dat?*. In a search of the English Brown corpus of written texts, Verhagen finds that 10 out of 11 examples of long-distance filler-gap constructions involved the verb *think*; in a search of a Dutch newspaper 34 out of 43 long-distance filler-gap constructions likewise involved the verb *denken* ('think').

Dabrowska (2004: this issue) reports sentence judgment studies in which she compared judgments on instances of the *WH do you think/say S?* formula with variations of the formula. Her study (this issue) demonstrates that questions of the form *WH do you think S?* are judged to be more grammatical than questions that instead involve the auxiliaries (*will* or *would*) or a different verb (*suspect, claim, swear, believe*) or that include an overt complementizer *that*.[3] See also Poulsen (2006) for similar findings for the verb *denken* ('think') in Dutch. One might quibble with certain aspects of Dabrowska's study; for example, half of the questions used as stimuli involved the verbs *think* or *say*, and it is possible that the repetition led subjects to give those instances higher ratings due to a general fluency effect (see e.g., Jacoby et al. 1989). In addition, we know that strings that contain more frequent words tend to be judged as more acceptable, all other things being equal (Ambridge et al. forthc. b; Featherston 2005; Keller 2000; Kempen and Harbusch 2003, 2004; Schuetze 1996); yet the high frequency *do* was compared with the low frequency *would*, and the high frequency *think* was compared with lower frequency verbs. Nonetheless, simply given the high frequency of *WH do you think S?* and *WH do you say S?* it seems reasonable to accept that these templates may be stored, as Dabrowska, Verhagen and Poulsen suggest.

Both Dabrowska (this issue) and Verhagen (2006) go farther, however, and argue that other instances of long distance dependency questions are judged by *analogy* to a fixed high-frequency formula, *WH do you think S?*. Verhagen (2006), for example, suggests that "Instances that do not conform to [the formulaic question], can be seen as analogical extensions

3. Dabrowska (this issue) finds no significant effect for changing the second person subject, you to a proper name, and the auxiliary must agree with the subject, so the stored formulas may be the more general WH DO NP think S? and WH DO NP say S?

from this prototype.... invented sentences exhibiting "long distance WH-movement" will be worse, the more they deviate from the prototype". Dabrowska (this issue) likewise suggests that in order to produce questions such as *What does she hope she'll get?* i.e., questions that do not fit the stored *WH do you think S?* template, speakers must adapt the template, substituting *she* for *you*, *hope* for *think*, and *does* for *do*. Bybee (2007) interprets usage-based theories to claim that grammaticality = familiarity, with general semantic or pragmatic constraints playing little role. She states, "Under the usage-based notion that lack of grammaticality is lack of familiarity, the oddness of these sentences [island violations] can be said to be in part due to the fact that one rarely hears such combinations of structures" (2007: 695).

If this view were extended to all constructions and combinations of constructions, it might be suggested that all of our knowledge of grammar is essentially item-based. What appear to be generalizations or novel combinations of constructions, would on this view, simply be one-shot analogies on memorized formulaic expressions.

Few researchers have actually defended a *purely* exemplar based model of linguistic knowledge, as usage-based models are not normally interpreted in this way. In particular, usage-based models espoused by Langacker (1988), Tomasello (2003) and Goldberg (2006) emphasize that speakers form generalizations over instances as they record specific instance-based knowledge (see also Murphy 2002 for a similar view of non-linguistic categorization). Dabrowska (2004: this issue) and Verhagen (2006) in fact, likewise take a moderate position, allowing that generalizations are often formed for constructions that are exemplified by a wide variety of examples in the input. Dabrowska has argued, for example in the case of other constructions, that "early usage is highly stereotypical and ... development proceeds from invariant formulas through increasingly general formulaic frames *to abstract templates*" (2004: 200, emphasis added). Verhagen (2006) also notes that higher type frequency of examples will lead to more abstract representations (see also Bybee 1985, 1995).

Still the question of whether we generalize beyond the exemplars is highly relevant to the present case in which the vast majority of attested examples instantiate only one or two relatively concrete types. We may grant that these types, namely the formulas *WH do you think S?* and *WH do you say S?* are likely to be stored, given their high frequency and the judgment data collected by Dabrowska (this issue). The issue raised by Dabrowska, Verhagen and Poulson's work is: is this *all* speakers have? Or, instead, is there evidence for a more abstract generalization about the function of long distance dependency constructions that enables

Table 2. *Islands involve non-assserted (here presupposed) information*

Complex NPs 1. She didn't see the report that was about him.	→	The report was about him.
Sentential subjects 2. That she knew it didn't bother him.	→	She knew it
3. She didn't leave the movie after they ate it	→	They ate it.
4. She didn't realize that he saw the roses.	→	He saw the roses.

us to combine the clausal complement and question constructions on the fly?

1.4. *Backgrounded Constructions are Islands (BCI) account*

Several researchers have argued that the constraints on filler-gap constructions are best accounted for in terms of certain discourse properties of the constructions involved. A fundamental insight of this perspective is the observation that the gap generally must fall within the potential focus domain of the sentence (Erteschik-Shir 1979; Erteschik-Shir 1998; Takami 1989; Deane 1991; Van Valin 1998; Van Valin and LaPolla 1997).[4] That is, the constituent in which the gap exists (i.e., the constituent containing the canonical position for the filler) must be within the part of the utterance that is asserted; it cannot be presupposed or otherwise "backgrounded." Presuppositions of a sentence are revealed by a classic negation test: presuppositions are implied by both the positive and negative form of a sentence. In accord with this observation, notice that all of the constructions in Table 1, with the exception of manner of speaking verb complements, convey presupposed information. This is indicated in Table 2: i.e., the negation of the sentences in Table 1, just like their positive counterparts, imply the propositional content expressed by

4. Van Valin's (1995) account suggests that the potential focus domain is defined structurally: that all direct daughters of direct daughters of the illocutionary force operator are within the potential focus domain. This account, like the subjacency account above, requires an appeal to other factors to explain the fact that the complements of manner of speaking and factive verbs are not fully acceptable since they are within his structurally defined potential focus domain (being direct daughters of direct daughters of the illocutionary force operator). A Gricean explanation has been offered for manner of speaking verbs (Van Valin 1997), but complements of factive verbs are predicted to be acceptable. In its favor, the "direct-daughters" proposal is aimed at predicting which constructions are non-backgrounded so that each construction need not be investigated on a case-by-case basis.

the island. Thus these island constructions do not express the assertion of a sentence: they are not part of the focus domain.[5]

Presupposition is a special case of non-assertion: what is presupposed is taken for granted by both the positive and negated version of a sentence. Another type of non-assertion is also revealed by the negation test, but is distinct from presupposition in that neither the embedded proposition nor its negation is implied by either the positive *or* the negated form of the sentence. Complements of manner-of-speaking verbs involve this type of non-assertion:

(18) She shouted that he left.
↛ He (didn't) left

(19) She didn't shout that he left.
↛ He (didn't) leave.

That is, normally a manner of speaking verb is only used when the manner of speaking and not the content of the complement clause is the main assertion of the clause:

(20) She didn't mumble that he left.
Natural interpretation: She didn't *mumble* the content.

Notice that in a context in which the manner of speaking can be taken for granted, the complement clause can be interpreted as asserted. For example, in a game of whisper-down-the-alley, main clause negation can be interpreted as negating the lower clause:

(21) I didn't whisper that the horse was green.
Natural interpretation: That the horse was green is not what I whispered. (e.g., I whispered that the house was clean)

As predicted by the information structure account, in this context, a gap within the complement clause is much improved:

(22) What did you whisper that the house was?

Thus we see that when the complements of manner-of-speaking verbs are not within the focus domain (i.e., not construed to convey the main assertion of a sentence), they are islands to extraction. In special contexts where they are construed to be within the focus domain, their island

5. In interpreting sentential negation, care must be taken not to place focal stress on any constituent. Contrastive or metalinguistic negation can negate content expressed within islands, but then this type of negation can be used to negate anything at all, including pronunciation or choice of lexical items, (*She didn't realize that he saw the ROSES, she realized that he saw CARNATIONS*!).

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 357)

status is noticeably mitigated. Thus the notion of "potential focus domain" is clearly relevant to island constraints, as many have noted for a long time (see references above).

At the same time, the potential focus domain does not capture the relevant facts perfectly. Subject complements are not within the focus domain, as they (or their existence) are presupposed:

(23) The king of France is bald.
→ There is a king of France.

(24) The king of France isn't bald.
→ There is a king of France.

And yet the entire subject argument is available for questioning:

(25) Who is bald?

The subject argument is not within the focus domain,[6] but it plays a special role in the information structure of a sentence in that it generally serves as the primary topic. In order to allow for the fact that (entire) subject arguments are available to serve as gaps, despite their not being within the focus domain of a sentence, Goldberg (2006: 135) formulates the discourse generalization as follows:

§ *Backgrounded* constituents may not serve as gaps in filler-gap constructions.
(Backgrounded constructions are islands: BCI)

Backgrounded constituents are defined as constituents that are neither the primary topic nor part of the focus domain of a sentence. Elements *within* clausal subjects are backgrounded in that they are not themselves the primary topic, nor are they part of the focus domain. Relative clauses, noun complements, presupposed adjuncts, parentheticals, and active ditransitive recipients are also not part of the focus domain of the clause and are therefore backgrounded (cf. Goldberg 2006). In this way, the account correctly predicts that a wide range of constructions should all be islands to long-distance dependency relations.

The restriction on backgrounded constructions is motivated by the function of the constructions involved. Elements involved in unbounded dependencies are positioned in discourse-prominent slots. It is pragmatically anomalous to treat an element as at once backgrounded and discourse-prominent.

6. Subject arguments may be within the focus domain in a limited type of sentence-focus construction (Lambrecht 1994). This construction requires special sentence accent on the subject argument and occurs with a restricted set of mostly intransitive verbs.

We have seen that the BCI predicts complements of factive verbs should be islands, since, by definition, the complements of factive verbs are presupposed and are therefore backgrounded. The complements of manner-of-speaking verbs are also predicted to be islands except in special contexts in which the manner is taken for granted. But as noted above the judgments of illformedness in these cases are somewhat subtle. While factive verbs more strongly presuppose the content of their complement clauses, it is not obvious that they are stronger islands than manner-of-speaking verbs, though this is what the BCI hypothesis predicts. Complements of semantically ''light'' bridge verbs (e.g., *say, think*) are predicted not to be islands, as these ''neutral'' verbs are generally used to introduce a complement clause containing the foregrounded information.

2. Testing the hypotheses

In this paper we set out to investigate the following questions: a) Do judgments relating to the negation test correlate with judgments concerning island status as the BCI account predicts? b) Do judgments concerning island status correlate with similarity of the main verbs involved to the verbs *think* and *say* as the direct-analogy proposal would predict?

We decided to restrict our investigation to one particular filler gap construction: long-distance WH-extraction from clausal complements. This allowed us to control for overall sentence length and complexity, as ratings were obtained for different verbs in exactly the same syntactic pattern. Four verbs were chosen from each of three classes of clausal-complement-taking verbs:[7]

a. factive verbs (realize, remember, notice, know)
b. manner-of-speaking verbs (whisper, stammer, mumble, mutter)
c. ''bridge verbs'' (say, decide, think, believe)

2.1. *Difference scores*

As described in detail in the methods section, we collected acceptability ratings for both WH-questions and the corresponding declarative statements. We used as our measure of acceptability of the WH-question a

7. We originally additionally included four whether-complement taking verbs, but these are treated as fillers in the analysis that follows. See Ambridge and Goldberg (forthc. a) for an analysis of whether as being intermediate between a complementizer and a WH-word.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 359)

difference score (or *dispreference for question-form score*) calculated by subtracting the rating for each WH-question from the rating for the corresponding declarative statement, averaging across all subjects for each item. For example, the number assigned to measure the dispreference for "extraction" in *Who did Pat stammer that she liked?* was arrived at by subtracting subjects' rating of this sentence from their rating of the corresponding declarative sentence, *Sara stammered that she liked Dominic*. This allows us to control for any general (dis)preferences that participants might have for particular VERB+COMP combinations. Such (dis)preferences might be expected to occur on the basis of simple frequency (e.g., sentences containing *say that* might be rated as more acceptable than sentences containing *stammer that*, regardless of whether they are interrogative or declarative) and/or (perhaps relatedly) the extent to which certain verbs felicitously introduce complement clauses (again in both declaratives and interrogatives). Indeed, in the present study, for example, declarative sentences of the form *NP said that S* received a mean rating of 5.9 out of 7, while sentences of the form *NP stammered that S* received a mean rating of 4.7. The finding that subjects give lower ratings of acceptability to sentences containing low frequency strings when other factors are held constant is well attested in the literature (see references cited earlier). Using these difference scores ensures that our dependent measure reflects the extent to which participants consider particular WH-extraction questions to be ungrammatical, controlling for the frequency of particular lexical strings. The higher the difference score, the higher the dispreference for the WH-question form (i.e., the higher the difference score, the stronger the island to extraction).

2.2. *Negation test*

A central goal of the study was to investigate whether the extent to which a complement clause is backgrounded correlates with its resistance to WH-extraction. As a measure of backgrounding of the complement clause, the negation test was used. The degree to which a clause C is considered backgrounded varies inversely with the extent to which main clause negation implies that C itself is negated. To determine scores on the negation test, we simply asked native speakers to judge the extent to which main clause negation implied that the subordinate clause was negated. For example, subjects judged the extent to which sentences like that in (26) implied (27) on a seven point scale:

(26) She didn't think that he left.
(27) He didn't leave.

Clearly (26) does not strictly entail (27), but it does imply it to some extent (as the judgments collected confirm). The negation test has the virtue of being a well-motivated, objective and independent measure. This test is intended to predict what is in the focus domain generally. For present purposes, the BCI predicts a correlation between the negation test and acceptability of the long distance dependencies, at least to the extent to which negation test judgments differ for particular verbs.

2.3. *Similarity judgments*

In order to determine whether semantic analogy to the verbs *think* or *say* play a role in acceptability judgments, we used both human and automatic calculations of semantic similarity. For the human judgment data, we created a second questionnaire to investigate verbs' similarity to *think* and *say*. For the automated calculation, we used Latent Semantic Analysis (Deerwester et al. 1990). The similarity judgments are discussed in section 6.

3. Predictions

3.1. *Predictions of the BCI hypothesis*

To recap, the BCI hypothesis predicts that the greater the extent to which sentential negation implies negation of the complement clause, the lesser the extent to which the complement clause is backgrounded, and hence the weaker the island. That is, the higher the negation-test score, the higher the predicted acceptability of the related WH-question, and the lower the difference score. Thus the BCI hypothesis predicts a significant negative correlation between negation-test and difference scores.

3.2. *Predictions of subjacency account*

A purely syntactic subjacency account would expect all structurally identical sentences to behave identically, and thus would predict no systematic differences across semantic verb classes. The proposals to treat complements of manner of speaking verbs as adjuncts and complements of factive verbs as part of complex NPs were argued to be problematic. However, if either of these analyses is correct it would predict that the constituent in question is an island to extraction. It is well-known that island status is somewhat variable, but no particular gradience of judgments is predicted on this account. That is, there is no reason to expect that grammaticality judgments should correlate in any systematic way with judgments on the negation test.

3.3. *Predictions of a direct analogy account*

Another possibility is that acceptability judgments (difference scores) are based on semantic similarity to an entrenched fixed formula involving the verb *think* or *say* (*WH do you think/say S?*). Dabrowska (this issue) found that judgments on questions involving the second person subject *you* were not significantly different from those with a proper name, so we might generalize the template to *WH DO NP think/say S?* where "DO" is capitalized to indicate that its form is determined by agreement with the subject argument. Our stimuli all contain past tense *did* and not *do* or *does*; this difference from the fixed formula is controlled for across items. Our stimuli all contain the complementizer *that* so this difference from the fixed formula is also controlled for across items. The key difference among our items is the main verb involved. The direct-analogy account would thus seem to predict that there should be a negative correlation between difference scores and scores of similarity of the main verbs involved to *think* or *say*: the more similar a verb is to *think* or *say*, the less the difference there should be between the acceptability of a question and the acceptability of its corresponding declarative.

4. Questionnaire #1: acceptability ratings and negation test

The first questionnaire collected acceptability judgments and judgments on a negation test. Similarity judgments were collected in a separate questionnaire (see section 6).

4.1. *Method*

4.1.1. *Participants.* Participants who filled out the acceptability/negation-test questionnaire were 71 naïve undergraduate and graduate students from Princeton University (mean age 19;6), all of whom were monolingual English speakers. None of the participants were linguistics majors and few if any had any background in linguistics. Participants received $5 for their participation during a questionnaire day.

4.1.2. *Design.* For each of twelve verbs, each participant rated the grammatical acceptability of a WH-question and a declarative statement —both containing a complement clause, and performed a negation-test-judgment task (see Materials section). The verb(class) was manipulated as a within-subjects factor with 12 levels for a correlation analysis, and three levels (*factive, manner of speaking, bridge*; with four verbs in each class) for a factorial analysis. Counterbalance-version (six different

versions of the questionnaire were used) was manipulated as a between-subjects factor.

4.1.3. *Materials.* Each participant completed a two-part questionnaire; the first part consisted of judgments of grammatical acceptability for WH-questions and declarative statements; the second part consisted of judgments about the extent to which main clause negation implied negation of the complement clause.

Acceptability judgments of WH-questions featuring WH-extraction from a clausal complement clause as in (A) were collected:

A) What did [NP1] [VERB1] [[that] [NP2] [VERB2]]?
(e.g., *What$_i$ did Jess think that Dan liked t_i*?)

VERB1 was one of the 12 experimental verbs: *realize, remember, notice, know; whisper, stammer, mumble, mutter; say, decide, think, believe*. NP1 and NP2 were one of 12 female or 12 male proper names respectively, while VERB2 was the past tense of one of 12 transitive verbs (*ate, bought, built, drew, fixed, found, knew, liked, made, needed, opened, pulled, read, threw, took, wanted*).

Six different versions of the questionnaire were created. For each version, sentences were generated at random using the template in (A).[8]

Acceptability judgments of declarative statements of the form given in (B) were also collected:[9]

B) [NP1] [VERB1] [that] [[NP2] [VERB2+APPROPRIATE NP]]
(e.g., *Danielle thought that Jason liked the cake*)

8. The actual sentence for each of the 12 experimental verbs (though not the structure of the sentence) differed across all six versions. For example, the experimental verb realize occurred in the sentence What did Ella realize that Adam threw? in Version 1, What did Trinity realize that Andy drew? in Version 2, and so on. This was to guard against the possibility of our findings being distorted by item effects.
9. Again, VERB1 was one of the 12 experimental verbs (this time in past tense form). As for questions of the form in (A), the declarative statements were generated at random using this template, and differed across the six versions of the questionnaire. VERB2 was selected from the same list of 12 verbs used for the questions, each paired with an appropriate NP (*ate the chips, bought the groceries, drew the picture, fixed the computer, found the keys, knew the secret, made the dinner, needed the map, pulled the car, read the book, threw the ball, wanted the chocolate*). NP1 and NP2 were selected from two further lists of 12 female and 12 male names (i.e., each name never appeared more than once throughout the study). This was to avoid explicitly highlighting to the subjects the formal relationship between each WH-extraction question and its equivalent declarative.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 363)

For each of the six questionnaire versions, the 24 items in part one of the questionnaire—12 WH-questions and 12 declarative statements—were presented in a different pseudo-random order, with the stipulation that no two verbs from the same verb class (*factive, manner of speaking, bridge*) were presented consecutively.

Negation test judgments. The second part of the questionnaire consisted of negation test judgments that were designed to indicate the extent to which sentential negation was interpreted as implying negation of (i.e., having scope over) the clausal complement. Each negated complex sentence (e.g., *Maria didn't know that Ian liked the cake*) was paired with a negated simple sentence corresponding to the complement clause of the complex declarative (e.g., *Ian didn't like the cake*). For each of the six different questionnaire versions, the complex + simple negated declarative sentence pairs were presented in a different pseudo-random order, with the stipulation that no two pairs involving verbs from the same class (*factive, manner of speaking, bridge*) were to be presented consecutively. These items (see Sentence C below for an example) were created using an additional set of 12 female names (NP1s) and male names (NP2), along with the same lists of VERB1s and VERB2+APPROPRIATE NPs as in the declarative statements from Part 1:

C) [NP1] didn't [VERB1] [that] [NP2] [VERB2+APPR. NP] [NP2] didn't [VERB2+APPR. NP]
   e.g., *Maria didn't know that Ian liked the cake Ian didn't like the cake.*

Again, the items generated for each verb differed across each of the six different versions of the questionnaire with regard to the NPs used.

4.1.4. *Procedure.* Subjects completed the questionnaire in written form, and were given only printed instructions.

For Part 1 (judgments of grammatical acceptability), these instructions stated:

> Please rate each of the sentences below for how acceptable you find them. 7 = Perfect (completely acceptable), 1 = Terrible (completely unacceptable).
>
> Please indicate your response by drawing a circle around the appropriate number as shown in the examples below. Please judge the sentences only on how acceptable you find them (and not, for example, whether the event they describe is plausible or implausible, good or bad etc.). Acceptability is a sliding scale and not a yes/no judgment—people tend to differ in their judgments of how acceptable sentences are.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 364)

For Part 2 (negation test judgments), these instructions stated:

> Here, you will be given two statements. Your task is to decide the extent to which the first statement implies the second statement. Consider the example sentence pairs in A–C below:
>
> (A) *Bob left early. Bob didn't leave early*.
> The first statement strongly implies that the second statement is NOT true, so in this case you would circle the 1, as shown above.
>
> (B) *Bob left the party early. Bob left the party*.
> This time, the first statement strongly implies that the second statement IS true, so this time, you would circle the 7 as shown above.
>
> (C) *Bob might leave the party late. Bob left the party early*.
> This time, the first statement neither implies nor does not imply the second statement, so here you would circle the 4 as shown above.
>
> We are interested in what average people typically imply with their everyday statements. Bearing these examples in mind, please rate the pairs below for the extent to which the first statement implies that the second statement is true. That is, if you heard a person say [Statement 1], to what extent would you assume that they are implying [Statement 2].

5. Results and discussion

Difference scores, raw scores (ratings for questions and declaratives), and negation-test scores can be found in Table A1 (Appendix).

5.1. *Preliminary analysis*

A preliminary analysis of variance with mean difference scores (preference for declarative over WH-extraction question) as the dependent variable and verb-type (*factive, manner of speaking, bridge*) as a within-subjects variable was conducted to investigate the effect of counterbalance version. This variable was not associated with any significant main effects or interactions. Subsequent analyses therefore collapsed across all six different questionnaire versions. The data—raw scores, difference scores and negation test scores—were also checked for normality of distribution (for each verb individually, and collapsed into the three verb-type categories). Although data in some conditions displayed skew and kurtosis, all subsequent analyses yielded the same pattern of results with raw and (log) transformed data. We therefore report results for untransformed data only.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 365)

5.2. *Analyses of variance*

In order to investigate the role of verb classes, we conducted an analysis of variance for difference scores and negation test scores separately, at the level of verb classes (*factive, manner of speaking, bridge*). That is, for each subject, the difference score for—for example—factive verbs represents the mean of that subject's difference scores for *realize, remember, notice* and *know* (and the same for negation-test scores).

These analyses were conducted to investigate (a) whether subjects gave significantly higher ratings of grammatical acceptability (looking at difference scores) for certain classes of complement-taking verbs than others and (b) whether participants' negation-test judgments mirrored (i.e., predicted) these acceptability ratings. These data are shown in Figure 1 and Figure 2 respectively (and also in Table A1; see Appendix).

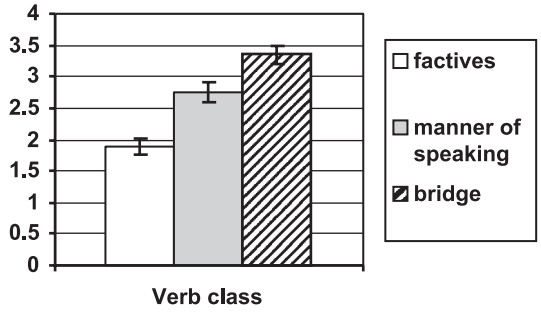

Figure 1. *Mean negation test scores. Higher scores indicate less backgrounding of the complement clause*

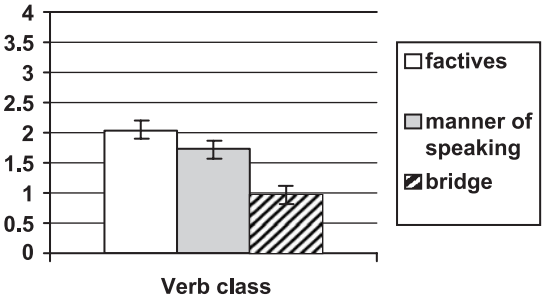

Figure 2. *Mean difference (dispreference-for-extraction-question) scores. Higher scores indicate greater ungrammaticality of the question form (relative to the corresponding declarative)*

As predicted by the BCI hypothesis, the increase in different scores is paralleled by a decrease in negation-test scores (recall that the BCI hypothesis predicts a *negative* correlation between our negation-test and difference-score measures).

A one-way within-subjects ANOVA with the independent variable of verb-type (*factive, manner of speaking, bridge*) and the dependent variable of difference score yielded a significant main effect of verb-type ($F_{(2,70)} = 27.01,\ p < 0.001,\ \eta_p^2 = 0.28$). Post hoc tests revealed that factive verbs yielded the strongest islands (i.e., highest difference scores; $M = 2.06$ places on the 7-point scale; $SE = 0.16$). Manner-of-speaking verbs ($M = 1.74$, $SE = 0.15$) yielded the next strongest islands, with (as their name implies) bridge verbs forming the weakest islands ($M = 0.97$, $SE = 0.13$). All comparisons were significant at $p < 0.001$ with the exception of that between factive and manner-of-speaking verbs, which was marginally significant at $p = 0.056$.

A one-way within-subjects ANOVA with the independent variable of verb-type (*factive, manner of speaking, bridge*) and the dependent variable of negation-test score also yielded a significant main effect of verb-type ($F_{(2,70)} = 49.27,\ p < 0.001,\ \eta_p^2 = 0.41$). Factive verbs yielded the lowest negation test score (i.e., highest backgrounding of the complement clause; $M = 1.90$, $SE = 0.13$), then manner-of-speaking verbs ($M = 2.75$, $SE = 0.15$), then bridge verbs ($M = 3.35$, $SE = 0.14$), with all comparisons significant at $p < 0.001$.

In summary, the results of these two ANOVAs provide considerable support for the BCI hypothesis. Factive verbs—which, as a class, are rated as strongly backgrounding the complement clause (as measured by the negation test)—form the least acceptable WH-extraction questions. Bridge verbs—which, as a class, are rated as only weakly backgrounding the complement clause—form the most acceptable WH-extraction questions, with manner of speaking verbs in-between the two.

In order to quantify the negative correlation between the difference scores and the judgments on the negation test, we additionally performed a correlation analysis on the data. The correlational analysis is affected by within-verb class correlations as well as correlations between verb classes, so it is a more sensitive measure.

5.3. *Correlation analyses*

We entered into the correlation analysis the mean negation-test score and the mean difference score, pooling across all subjects (see Lorch and Myers 1990). What our analysis lacks in power—having only 12 datapoints—it makes up for in reliability, as each point includes scores

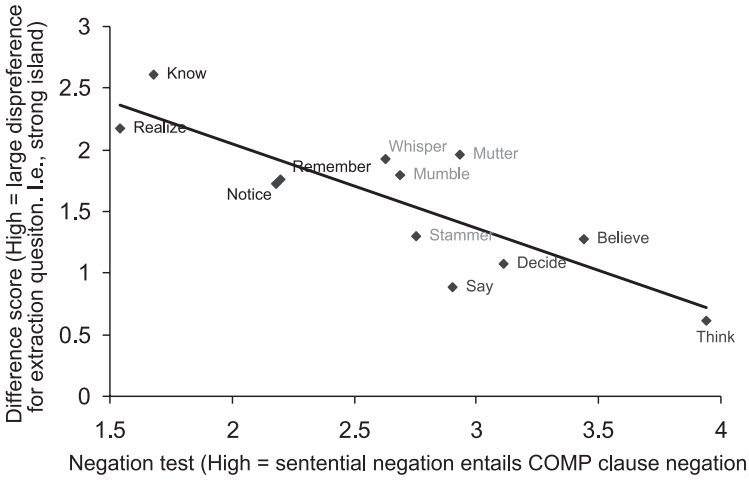

Figure 3. *Correlation between difference scores (dispreference for question scores) and negation test scores*

from 71 participants. A scatterplot of this correlation is shown in Figure 3.

This analysis revealed that the mean negation test score was a highly significant (negative) predictor of mean difference score ($r = -.83$, $p = 0.001$), accounting for over two thirds of the observed variance ($R^2 = 0.69$).[10] The correlation of |.83| is strikingly high, as perfect correlations (+/−1) are almost non-existent when distinct measures are used. Separate measures of *the same thing*, e.g., mean length of utterance (MLU) at 28 months, have been found to correlate in the .75–.80 range (Bates and Goodman 1997).

5.4. *Any role for subjacency?*

The subjacency account clearly does not predict the pattern of results found in the present study. In particular, subjacency does not predict any distinctions based on the semantic class of the verbs involved without

10. Mean negation test score was also a significant (positive) predictor of mean rating of acceptability for the extraction question ($r = 0.58$, $p < 0.05$), accounting for approximately one third of the observed variance ($R2 = 0.34$). Thus although, as we have argued, difference scores constitute a more appropriate measure of (un)acceptability than raw scores, our finding of a significant association between backgrounding and the acceptability of WH-extraction questions does not hinge on using difference scores.

stipulation. Manner of speaking complements and factive complements would require reanalysis as adjuncts or parts of complex NPs as outlined above in order to predict their relative ill-formedness *vis a vis* semantically light verbs. Such analyses would require independent support, of course, or risk being ad hoc; moreover, even if reanalysis into adjuncts and complex NPs is granted, it does not predict the strong correlation found between difference judgments and judgments on the negation test. Moreover, subjacency does not predict the fact that questions from complements of the verb *think* (and *say*) are judged to be particularly well-formed.

5.5. *'Think' and 'say' questions as stored formula*

As Figure 3 illustrates, the present study replicated Dabrowska's (2004: this issue) findings that WH-questions with *think* and *say* are rated as somewhat more acceptable than such questions with other verbs (*think* is significantly more acceptable than all other verbs besides *say* and *decide* at $p < 0.05$ by paired *t*-test; *say* is more acceptable than all other verbs except *think, believe, decide* and *stammer*). At the same time, the grammaticality judgments do not provide unambiguous evidence for formulaic status since the semantic properties of *think* and *say* predict that they should be favored. To demonstrate that speakers judge WH-extraction questions with *think* and *say* to be more acceptable than would be predicted given their semantics, it would be necessary to show that scores for these items fall well below the regression line. Generally a difference of 1.96 standard deviations is accepted as indicating outlier status and neither *say* nor *think* meet this criterion. Although *say* is 1.62 standard deviations below the regression line, and thus farther from the regression line than most of the other verbs, it does not meet this criterion for outlier status; neither is it the closest to being classified as an outlier (*mutter* is judged 1.65 SDs worse than would be predicted given its negation test score). Moreover, acceptability of the WH-extraction question is predicted better by the negation test for *think* than for any other verb (at only −0.28 SDs below the regression line).

The BCI generalization goes some way toward explaining why the same verbs, "think" and "say," are more likely to appear in long-distance dependency constructions than other verbs cross-linguistically: their semantics motivates their discourse properties which in turn motivate their distribution (recall that e.g., Dutch *denken* "think" shows a tendency to be used frequently in filler-gap constructions (Verhagen 2006); and cf. discussion of Polish "say" below). The idea that *think* is used with special discourse properties is buttressed by the idea that clauses with the main

verb *think* are often cited as in some sense "monoclausal" (cf. also Lakoff 1969; Thompson 2002; Verhagen 2006). An indication that *say* is likewise often used to foreground the information in the complement clause comes form the fact that the verb *say* has been known to grammaticalize into a complementizer (Haspelmath 1989).

Thus although we agree that the forms *WH DO NP think S?* and *WH DO NP say S?* are *likely* to be stored (due to their high frequency), the present study does not provide evidence that this is *necessarily* the case, as their well-formedness may be due to their semantics. To demonstrate that WH-extraction questions with *think* and *say* are necessarily stored as templates, one might turn to on-line comprehension time measures which may be more likely to reveal formulaic status than acceptability judgments (cf. Wonnacott et al. forthc.). The following section investigates the stronger claim that these high-frequency formulas are used as the basis of direct semantic analogy when other WH-questions with the same form are at issue.

6. Questionnaire #2: Semantic similarity

As noted at the outset, the direct-analogy proposal claims that the questions *WH DO NP think S?* and *WH DO NP say S?* constitute semantic prototypes, and that the grammatical acceptability of other such questions may vary as a function of their semantic similarity to these prototypes. In order to test this possibility, we investigated whether semantic similarity of each verb to *think* (or *say*) accounted for any of the observed variance in difference scores.

6.1. *Method*

6.1.1. *Participants.* 12 naïve undergraduate and graduate students (11 from Princeton and one from the University of Liverpool) (mean age 22.5) filled out semantic similarity questionnaires. None of them had taken part in the first study and as before, none of the participants were linguistics majors, and few if any had any background in linguistics. Participants received $7 each.

6.1.2. *Design.* Participants rated the semantic similarity of *say* and *think* (and—as a control—four other verbs used in the main study) to each of the 11 remaining verbs (see Materials section). The control verbs allow us to determine whether semantic similarity to *think*/*say* in particular predicts the difference scores from the first study better than semantic

similarity to an arbitrary verb that is not claimed to form part of a semantic template (e.g., *remember*). Each subject received one of six differently ordered versions of the questionnaire in order to guard against possible order-effects.

6.1.3. *Materials.* In order to give semantic similarity the strongest chance of predicting the acceptability of question forms, we asked speakers to judge the similarity of the verbs as they appeared in questions. We used yes/no questions because judgments on WH-questions would have been confounded by variation in acceptability, which may have influenced speakers' similarity ratings in unforeseen ways. Participants filled out a questionnaire containing items such as the following:

> How (dis)similar are the following verbs to *think*, in the context
>
> A. Did you *think* that Mary needed the map?

Did you decide that Mary needed the map?	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar
Did you say that Mary needed the map?	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar
Did you whisper that Mary needed the map?	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar

The verbs *think, say, remember, notice, stammer* and *mumble* were used in target questions as *think* is in the question in A above. For each target verb, similarity ratings were requested for each of the other 11 verbs (*realize, remember, know, whisper, mutter, decide*, and *believe* in addition to those used in the target sentences).

6.1.4. *Procedure.* Subjects completed the questionnaire in written form, and were given only printed instructions:

> Your task in this study is to rate verbs for how similar in meaning they are to another verb (as it is used in a particular sentence). For example, consider the sentence
>
> John *saw* the man.
>
> You might decide that—in this context—"spotted" means something very similar to "saw", in which case you would circle the 7 as shown below:

John *spotted* the man.	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar

You might also decide that—in this context—"kicked" means something entirely different to "saw", in which case you would circle the 1, as shown below:

John *kicked* the man.	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar

Finally, you might decide that—in this context—the meaning of "watched" is not *very* similar to that of "saw", but it is not *very* different either, in which case you would circle the 5 as shown below:

John *watched* the man.	Meanings are very different	1 2 3 4 5 6 7	Meanings are very similar

6.2. *Results and discussion*

The second questionnaire aimed to determine whether there was evidence for the idea that *think* or *say* WH-extraction questions were used as the basis for an analogy when judging the well-formedness of such questions with other main verbs. We therefore entered into a correlation analysis, for each verb (except *think* itself), the score representing the semantic similarity of this verb to *think* (predictor variable) and the mean difference score from the first study (outcome variable). A separate correlation analysis was performed for semantic similarity to *say*, and also to each of the four "control" verbs in the same way. The mean semantic-similarity to *think* (and *say*) scores are shown in Table A1 (Appendix).

The semantic-similarity judgment data failed to show a significant correlation with the judgment data for well-formedness of questions (i.e., difference scores). The correlations did not approach significance for similarity to either *think* ($r = 0.08$, $p = 0.79$) or *say* ($r = 0.17$, $p = 0.62$), (or, indeed, for any of the four control verbs: *remember, notice, stammer* or *mumble*). Indeed, the small and non-significant correlations for *think* and *say* were in the opposite direction to that predicted by the analogy account.

Relatively few subjects (12) were involved because preliminary analysis showed that judgments were highly reliable across participants. Each individual participant's judgments were significantly correlated with the

mean scores collapsing across all participants at $p < 0.01$. Note also that because we used mean scores pooled across participants, the power of the statistical test is unaffected by sample size. The validity of our analysis (as well as its power to detect effects) is demonstrated by systematic finding of significant correlations between similarity scores. For example, similarity-to-*think* scores were significantly (negatively) correlated with similarity-to-*say* scores $r = -0.741$, $p < .02$). In fact, judgments of similarity to all target verbs (*say, think, remember, notice, stammer* and *mumble*) were intercorrelated at $p < 0.05$ or better, with one exception (similarity-to-*notice* and similarity-to-*stammer*: $r = -.586$, $p = 0.075$).

Perhaps participants were basing their similarity judgments on some sort of conscious strategy that was not relevant to the implicit similarity judgments that might be used on the analogy proposal. To test for this possibility, we also calculated similarity scores using an on-line automatic similarity calculator, Latent Semantic Analysis (Deerwester et al. 1990).[11] As before, since a higher LSA score indicates greater semantic similarity to *think* (or *say*), and a lower difference score indicates a higher rating of grammatical acceptability, a negative correlation between LSA and mean difference score was predicted.

The analysis found that LSA semantic similarity of the verbs to *say* did not involve a significant correlation ($r = -0.02$, $p = 0.96$). Similarity to *think* was also not a significant predictor of mean difference score; in fact there was again a small non-significant correlation in the opposite direction to that predicted ($r = 0.11$, $p = 0.75$).

Another potential correlation we considered involved determining, for each verb, the maximum similiarity score to *either* the verb *think* or the verb *say*. That is, if we assume that two distinct formulas are stored, *WH DO NP say S?* and *WH DO NP think S?* then judgments may be determined by a comparison between a target verb and whichever formula it is semantically closest to. We therefore calculated the correlation between the difference scores and the array of scores determined by the following formula:

Max (similarity-of-verb$_i$-to-*say*, similarity-of-verb$_i$-to-*think*)
for i $\in$ {*realize, remember, notice, know, whisper, stammer, mumble, mutter, decide, believe*}.

However, neither the judgment scores of similarity nor the LSA similarity scores correlated significantly with the difference scores by this

11. See http://lsa.colorado.edu/ (texts denoted "General Reading up to 1st Year College").

measure either ($r = -0.17$, $p = 0.64$; $r = 0.45$, $p = 0.19$, respectively). The correlation with LSA scores is of a fair size ($r = 0.45$), but it is in the *opposite* direction to that predicted by the direct analogy account. Recall that difference scores are smaller to the extent that the question form is relatively well-formed. If judgments were based on semantic analogies to the fixed formulas, there should be a negative, not a positive correlation.

Thus, whichever way it is analyzed, by similarity to *say* or to *think* or to their combination, and according to either human judgment data or to the automatic LSA similarity calculation, semantic similarity to *say* or *tell* is a poor predictor of judgment data. The direct-semantic analogy proposal fails to account for the data.

7. Conclusion

In conclusion, the BCI hypothesis (or an information structure account more generally) has been shown to be an excellent predictor of the island status of clausal complements. Participants' negation-test judgments were able to predict over two-thirds of the variance associated with their dispreference-for-WH-extraction-question scores. As this correlation is also not expected nor easily explained on a purely syntactic account, this finding lends strong support to the idea that the discourse function of the constructions involved plays a critical role in island phenomena.

It is beyond the scope of this paper to provide a thorough comparison of the BCI and subjacency, but there are many other generalizations that the BCI accounts for without additional stipulation that subjacency does not (see Goldberg 2006). The first two predictions were considered in the present study.

1. Complements of manner-of-speaking verbs and factive verbs are islands.
2. Grammaticality judgments should correlate with the degree of ''backgroundedness'', when length and complexity are held constant.
3. Direct replies are sensitive to islands (Morgan 1975).
4. Exclamative ah! is sensitive to islands (James 1972).
5. The active recipient argument of ditransitive, as a secondary topic, resists being a gap, while the passive recipient argument of a ditransitive, as a primary topic, is free to be a gap.
6. Presentational relative clauses are not always islands.
7. Definite relative clauses are stronger islands than indefinite relative clauses.
8. Parentheticals are islands.

There is ample evidence that general processing constraints play a role in island violations (and their amelioration) (cf. e.g., Ellefson and Christiansen 2000; Gibson 1998; Kluender and Kutas 1993). Several factors including length, definiteness, complexity, and interference effects (involving similar referents between filler and gap) have been shown to play a role. As the present experiment controls for these factors, we can see that information structure constraints play an independent role in addition to effects of processing.

Judgments on filler-gap constructions involving the complement clause of the main verb *think* (and *say*) were judged to be significantly more acceptable than those involving most other main verbs, as Dabrowska (2004: this issue) has also found. The BCI hypothesis actually predicts that these verbs should be preferred on semantic grounds—the acceptability judgments correlate well with the negation test scores—so other data are needed to confirm that a fixed formula is stored (but, again, we take the idea that such formula are stored to be quite plausible).

On the other hand, the possibility that all filler-gap expressions involving complement clauses are judged by direct analogy to the formulaic expression with *think* or *say* was not supported by the data. Neither the human similarity judgment scores nor the automated LSA similarity measure correlated with the acceptability data. This finding argues against a strong version of item-based grammar in which acceptability judgments are necessarily determined by one-shot analogies to well-learned formulaic patterns.

In general, we must be careful when appealing to frequency in the input data as an explanation for linguistic generalizations. The explanation may be question-begging unless an account is offered as to why there should be cross-linguistic generalizations about the nature of the input, as there are, at least to some extent, in the case of island constraints. We must ask, why is the input the way it is? An account that appeals to information structure provides an answer to this question: speakers avoid combining constructions that would place conflicting constraints on a constituent, such as requiring it to be at once backgrounded *and* discourse-prominent.

At the same time, certain cross-linguistic differences do exist. As noted above (n. 4), Russian allows gaps only in main clauses, whereas Italian appears to allow long distance dependencies somewhat more freely than English. Insofar as backgroundedness is a matter of degree, languages appear to select different cut-off points in how backgrounded a constituent may be while containing a gap (cf. Erteschik-Shir 1973; Fodor 1991 for similar suggestions). Languages differ as to the location of the cut off point, but all languages seem to prefer extraction out of non-backgrounded constituents.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 375)

One further intriguing piece of evidence that suggests that conventionality (item-based learning) plays a role in addition to the information structure generalization comes from the fact that there are some cross-linguistic differences in which verbs within the class of ''bridge verbs'' are most likely to allow extraction from their complement clause. According to Cichocki (1983), the Polish verb *mówić* ('say') allows extraction from its finite complement clause while other verbs, including *myśleć* ('think'), do not.[12] At the same time, there are certain intriguing differences in how Polish *myśleć* and English *think* are used that deserve further exploration.[13] In any case, *say*, like *think*, is a light verb which allows its complement clause to be foregrounded (as evidenced by the present negation test scores). Thus, while the relative difference between Polish's ''think'' and ''say'' is not necessarily predicted by the BCI account, the following more general prediction is made: we do not expect to find any language in which a factive verb or a manner-of-speaking verb is more likely to allow extraction from its complement clause than a light verb of thinking or saying.

There is a vast and growing amount of evidence that speakers are aware of detailed statistical patterns in the input. We in no way wish to deny this. Certainly, speakers' inventories of constructions are learned by generalizing over instances, and the generalizations are often statistical in nature. The effects of statistics in the input are also clearly relevant to language processing (cf. e.g., other papers in this issue).

12. We thank Ewa Dabrowska and Blazej Galkowski for confirming the preference for extraction with Polish ''say'' over ''think,'' although Dabrowska notes that even extraction out of ''say'' complements is not fully grammatical in Polish (p.c. 20 March 2007 and 2 May 2007, respectively).

13. Galkowski (p.c. 2 May 2007) observes that Polish myśleć cannot be used as a hedge to assert the content of the subordinate clause the way English ''think'' can be, when there is main clause negation. He suggests that a more elaborate context in which the thought processes of the subject argument are at issue is required for the following type of example.

Nie	myślę	że	(on)	zjadł	tego	hamburgera.
Not	think-1sg	that	he	eat-3sg-past	this/the	hamburger

'I don't think he ate the hamburger'

For example, Galkowski offers the following context: ''[My grandpa with Alzheimer's can't be trusted to eat the food I leave for him. So when I see the plate is empty, I don't think he ate the hamburger. I'd rather look for it under his bed.] So the emphasis is on thinking''—BG (p.c. 2 May 2007). Insofar as the focus is on the main verb ''think'' and not the complement clause, the information structure account would predict that extraction from the complement clause should be dispreferred, as it is.

(AutoPDF V7 27/5/08 11:46) WDG (148×225mm) TimesM J-1972 COGL, 19:3 PMU: WSL(W) 23/05/2008 pp. 349–382 1972_19-3_01 (p. 376)

Yet constructions are combined to form actual expressions, and it seems unlikely that every possible combination of constructions is somehow stored in advance. The present studies undermine the position that the felicity of combination is always determined by semantic comparison with a relatively concrete, fixed formula. They also undermine any purely structural account such as subjacency. Rather, the current findings support a view of grammar in which speakers determine which constructions can be combined, at least in part, on the basis of the information structure properties of the constructions involved.

Received 21 May 2007
Revision received 21 December 2007

University of Liverpool, UK
Princeton University, USA

Appendix

Table A1 shows, for each verb (class), mean ratings of grammatical acceptability (and corresponding standard deviations) for questions and declarative sentences, difference scores, negation-test scores, and human/latent semantic analysis similarity-to-*think* scores).

Table A1. *Raw data*

Verb (class)	*Difference Score*		*Question*		*Declarative*		*Negation test score*		Judged similarity to *think-say*	LSA similarity to *think-say*
	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
Realize	2.17	1.75	4.13	1.76	6.30	1.14	1.54	1.12	4.58–3.00	0.49–0.41
Remember	1.76	2.25	4.27	1.94	6.03	1.36	2.20	1.74	2.67–2.42	0.55–0.50
Notice	1.72	1.70	4.45	1.77	6.17	1.40	2.18	1.72	3.08–2.25	0.38–0.30
Know	2.61	1.85	3.82	1.84	6.42	1.28	1.68	1.27	4.50–3.42	0.70–0.63
Whisper	1.92	1.96	3.99	1.81	5.90	1.57	2.63	1.55	1.58–4.75	0.28–0.33
Stammer	1.30	2.19	3.41	1.75	4.70	1.82	2.75	1.69	1.50–4.75	0.15–0.27
Mumble	1.79	2.10	3.96	1.78	5.75	1.48	2.69	1.65	1.42–4.67	0.11–0.20
Mutter	1.96	2.05	3.90	1.97	5.86	1.53	2.93	1.61	1.83–3.92	0.14–0.25
Say	0.89	2.06	5.04	2.00	5.93	1.66	2.90	1.48	2.08–N/A	0.62–N/A
Decide	1.07	1.85	4.73	1.82	5.80	1.59	3.11	1.59	4.33–2.42	0.34–0.28
Think	0.62	1.64	5.28	1.64	5.90	1.71	3.94	1.80	N/A–3.42	N/A–0.62
Believe	1.28	1.95	4.86	1.78	6.14	1.40	3.44	2.01	5.83–2.33	0.52–0.51
Factives	2.06	1.33	4.17	1.43	6.23	1.05	1.90	1.07		
Manner of speaking	1.74	1.25	3.81	1.30	5.55	1.20	2.75	1.26		
Bridge	0.96	1.13	4.98	1.42	5.94	1.21	3.35	1.21		

References

Ambridge, Ben and Goldberg, Adele E.
forthc. a Is *whether* a complementizer or a WH-word?

Ambridge, Ben, Pine, Julian M., Rowland, Carolyn F., and Chris R. Young,
forthc. b The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgments of argument-structure overgeneralization errors. *Cognition*.

Ambridge, Ben, Caroline F. Rowland, and Julian Pine
2007 Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science* 31(6),

Ambridge, Ben, Caroline F. Rowland, Anna L. Theakston, and Michael Tomasello
2006 Comparing different accounts of inversion errors in children's non-subject wh-questions: 'What experimental data can tell us?' *Journal of Child Language* 33(3), 519–557.

Baltin, Mark R.
1982 A landing site for movement rules. *Linguistic Inquiry* 13, 1–38.

Bates, Elizabeth and Judith Goodman
1997 On the inseparability of grammar and lexicon: evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processing* 12, 507–84.

Bybee, Joan L.
1985 *Morphology*. John Benjamins.

Bybee, Joan L.
1995 Regular morphology and the lexicon. *Language and Cognitive Processes* 10, 425–55.

Bybee, Joan L.
2007 Review of Goldberg's *Constructions at Work: the nature of generalization in language. Journal of Child Language.*

Chomsky, Noam
1973 'Conditions on Transformations', in S. R. Anderson and P. Kiparsky (eds) A Festschrift for Morris Halle, New York: Holt, Rinehart and Winston.

Cichocki, Wladyslaw
1983 Multiple Wh-questions in Polish: a two-comp analysis. *Toronto working papers in linguistics* 4, 53–71.

Dąbrowska, Ewa
2004 *Language, Mind and Brain*. Georgetown: Georgetown University Press.

Deane, Paul
1991 Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics* 2, 1–63.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harsham
1990 Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 391–407.

Ellefson, Michelle R. and Morten Christiansen.
2000 Subjacency Constraints without Universal Grammar: Evidence from Artificial Language Learning and Connectionist Modeling. *Proceedings of the 22nd Annual conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 645–650.

Erteschik-Shir Nomi
1973 On the nature of island constraints. Indiana University Linguistics Club.
Erteschik-Shir, Nomi
1979 Discourse contraints on dative movement. in S. Laberge and G. Sankoff (eds.) *Syntax and Semantics: Vol 12: Speech Acts.* New York: Academic Press, 441–467.
Erteschik-Shir, Nomi
1998 *The dynamics of focus structure*. Cambridge: Cambridge University Press.
Erteschik-Shir, Nomi and Shalom Lappin
1979 Dominance and the functional explanation of island phenomena. *Theoretical Linguistics* 6, 41–85.
Featherston, Sam
2005 Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* 115, 1525–1550.
Fodor, Janet D.
1991 Sentence processing and the mental grammar. In P. Sells, S. M. Shieber, and T. Wasow (Eds.), *Foundational Issues in Natural Language*. Cambridge, MA: MIT Press.
Freidin, Robert and Carlos Quicoli
1989 Zero-stimulation for parameter setting. *Behavioral and Brain Sciences* 12, 338–39.
Gibson, Edward
1998 Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76.
Goldberg, Adele E.
2006 *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.
Haspelmath, Martin
1989 From purposive to infinitive- a univeral path of grammaticalization. *Folia Linguistica Historica* 1–2, 287–310.
Jacoby, Larry L., Charlotte M. Kelley, and Jane Dywan
1989 Memory attributions. Varieties of memory and consciousness: essays in honour of ednel tulving, ed. by H. L. Roediger and F. I. M. Craik, Hillsdale, NJ: Erlbaum, 391–422.
James, Deborah
1972 Some aspects of the syntax and semantics of interjections. Paper presented at the 8th Regional Meeting of the Chicago Linguistic Society, Chicago.
Keller, Frank
2000 Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Unpublished PhD Thesis, University of Edinburgh.
Kempen, Gerard, and Karin Harbusch
2003 An artificial opposition between grammaticality and frequency: Comment on Bornkessel, Schlesewsky and Friederici (2002). *Cognition* 90, 205–210.
Kempen, Gerard, and Karin Harbusch
2004 The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In S. Kepser and M. Reis (eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton De Gruyter.
Kiparsky, Paul and Carol Kiparsky
1971 *Fact. Progress in Linguistics*, M. Bierwisch and K. Heidolph (eds.). The Hague: Mouton.

Kluender, Robert and Marta Kutas
1993 Subjacency as a Processing Phenomenon. *Language and Cognitive Processes* 8, 573–633.
Lakoff, Robin
1969 A syntactic argument for negative transportation. Proceedings of the Chicago Linguistic Society 15, 140–49.
Lambrecht, Knud
1994 *Information Structure and Sentence Form.* Cambridge: Cambridge University Press.
Langacker, Ronald W.
1988 A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*. Amsterdam: Benjamins, 127–161.
Lorch, Richard F. J., and Jerome L. Myers
1990 Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16(1), 149–157.
Morgan, Jerry L.
1975 Some interactions of syntax and pragmatics. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, Vol. 3: Speech Acts.* New York: Academic Press.
Murphy, Gregory L.
2004 *The Big Book of Concepts*. Cambridge, MA: MIT Press.
Newmeyer, Frederick J.
1991 Functional explanation in linguistics and the origins of language. *Language and Communication* 11(1–2), 3–28.
Pinker, Steven, and Paul Bloom
1990 Natural language and natural selection. *Behavioral and Brain Sciences* 13(4), 707–784.
Poulsen, Mads.
ms-2006 The time course of extraction constraints in Dutch.
Rizzi, Luigi.
1982 *Issues in Italian Syntax.* Dordrecht: Foris.
Ross, John R.
1967 Constraints on variables in syntax, MIT. Published as *Infinite Syntax!*. 1986. Norwood, NJ: Ablex Publishing Corporation: Ph.D.
Sag, Ivan A. and Janet D. Fodor
1994 Extraction without Traces. *WCCFL*, 13.
Schuetze, Carson T.
1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methology*. Chicago: University of Chicago Press.
Takami, Ken-ichi
1989 Prepositional stranding: arguments against syntactic analyses and an alternative functional explanation. *Lingua* 76, 299–335.
Thompson, Sandra A.
2002 "Object complements" and conversation towards a realistic account. *Studies in Language* 26, 125–64.
Tomasello, Michael
2003 *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.
Van Valin, Robert
1993 Synopsis of RRG. Advances in Role and Reference Grammar, ed. by Robert Van Valin: Benjamins.

Van Valin, Robert
1998 The acquisition of wh-questions and the mechanisisms of language acquisition. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Valin, Robert
1995 Toward a functionalist account of so-called 'extraction constraints'. In B. Divriendt et al. (Ed.), *Complex Structures: A Functionalist Perspective*. Berlin: Mouton de Gruyter, 29–60.

Van Valin, Robert, and Randy J. LaPolla
1997 *Syntax: structure, meaning and function.* Cambridge: Cambridge University Press.

Verhagen, Arie
2006 On subjectivity and ''long distance Wh-movement''. *Subjectification: Various paths to subjectivity*, ed. by A. Athanasiadou, Costas Canakis and Bert Cornillie. New York: Mouton de Gruyter, 323–46.

Wonnacott, Elizabeth, Elissa L. Newport, and Michael K. Tanenhaus
2007 Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*.