Questions with long-distance dependencies: 
A usage-based perspective

EWA DĄBROWSKA*

Abstract

Attested questions with long-distance dependencies (e.g., What do you think you’re doing?) tend to be quite stereotypical: the matrix clause usually consists of a WH word, the auxiliary do or did, the pronoun you, and the verb think or say, with no other elements; and they virtually never contain more than one subordinate clause. This has lead some researchers in the usage-based framework (Dąbrowska 2004; Verhagen 2005) to hypothesise that speakers’ knowledge about such constructions is best explained in terms of relatively specific, low level templates rather than general rules that apply “across the board”. The research reported here was designed to test this hypothesis and alternative hypotheses derived from rule-based theories.

Keywords: Usage-based model; long-distance dependencies; unbounded dependencies; acceptability judgment experiment; prototype effects.

1. Introduction

Questions and other constructions with long distance dependencies (henceforth LDDs) have played an important role in the development of syntactic theory, especially in the generative framework. Such structures

* This study was supported by the Arts and Humanities Research Council (grant number AH/F001924/1). I would like to thank Adele Goldberg, Marcin Szczerbiński and two anonymous referees for their comments on an earlier draft of this paper, and Mike Pincombe for help with data collection. Address for correspondence: School of English Literature, Language and Linguistics, University of Sheffield, Sheffield S10 2TN, United Kingdom. E-mail: e.dabrowska@shef.ac.uk.
are interesting because they exhibit a dependency between a “filler” in the main clause and a “gap” in a subordinate clause, as in example (1). The dependencies are frequently referred to as “unbounded”, as, in principle, there can be any number of clauses intervening between the filler and the gap, as illustrated by (1d) and (1e).

(1) a. What will John claim that you did __? (Culicover 1997: 184)
b. Which problem does John know (that) Mary solved __? (Ouhalla 1994: 72)
c. Whom do you believe that Lord Emsworth will invite __? (Haegeman 1991: 342)
d. Who did Mary hope that Tom would tell Bill that he should visit __? (Chomsky 1977: 74)
e. Which problem do you think (that) Jane believes (that) Bill claims (that) Mary solved __? (Ouhalla 1994: 71)

It is noteworthy, however, that attested questions with LDDs are very different from these constructed examples, as illustrated by the following examples from the spoken part of the British National Corpus:

(2) a. And how do you think you’d spell classical like do you like classical music? (FMG 725)
b. Why’d why do you think why do you think it is that there wasn’t that motivation? (FY8 201)
c. What is it and why do you think it looks like that? (JJS 882)
d. What do you think Brian’ll say? (KE1 256)
e. What did they say it meant? (KD0 622)

These “real life” LDD questions are much more stereotypical than the sentences in (1). The textbook examples contain a variety of matrix subjects and verbs and different auxiliaries; most of them also contain an overt complementizer and two involve a dependency over more than one intervening clause. In the corpus sentences, in contrast, the matrix subject is usually you, the matrix verb think or say, and the auxiliary do; there are no other elements in the matrix clause, no complementizer, and only one complement clause. In fact, almost 70 percent of the LDD questions with finite complement clauses in the spoken part of the BNC have the form WH do you think S-GAP? or WH did NP say S-GAP?, where S-GAP is a subordinate clause with a missing constituent. Most of the remaining questions are minimal variations on these patterns: that is to say, they contain a different matrix subject or a different verb or a different auxiliary or an additional element like an adverbial or complementizer. Only
6 percent depart from the prototype in more than one respect (Dąbrowska in press a; see also Dąbrowska 2004 and Verhagen 2005). This has lead some researchers in the usage-based framework (Dąbrowska 2004, Verhagen 2005) to hypothesise that speakers’ knowledge about such constructions is best explained in terms of relatively specific, low level templates—WH do you think S-GAP? and WH did you say S-GAP?—rather than in terms of abstract rules and principles of the type proposed by formal linguists (see, for example Cheng and Corver 2006; Chomsky 1977; Levine and Hukari 2006). Declaratives with verb complement clauses, in contrast, are much more varied—the main clauses take different subjects, auxiliaries and verbs, and often contain additional elements (Dąbrowska in press a; Verhagen 2005)—as a result of which language learners develop more general representations for this construction in addition to lexically specific templates for frequent combinations such as I think S, I don’t think S, I mean S. However, conclusions about speakers’ mental representations based on the fact that a particular structure is rare or not attested at all in a corpus are problematic, since this could be merely a result of sampling. Even with a large and balanced corpus, sentences which are perfectly compatible with speakers’ mental grammars may be unattested simply because they are pragmatically implausible. In short, while restricted patterns of usage are suggestive, they do not license strong conclusions about mental representation: the observational data need to be corroborated by experimental studies.

According to the usage-based proposals put forward by Dąbrowska and Verhagen, prototypical LDD questions are produced simply by inserting new material into the appropriate slots in a pre-existing template. Non-prototypical LDD questions such as What does she hope she’ll get? require additional work, since the speaker has to adapt the template—in this case, substitute she for you and hope for think, and modify the auxiliary so that it agrees with the subject. To be able to do this, the speaker would have to construct a proportional analogy such as the one in (3),

1. I am using the term “prototype” as it is usually used in linguistics: to refer to an idealized typical instance. Many natural categories are centred around a prototype, in the sense that other instances are assimilated in the category on the basis of their perceived similarity to it (cf. Lakoff 1987; Langacker 1987). The properties of prototypical instances are thus shared by most other members of the category.

2. The term “construction” is used in this paper to refer to any grammatical pattern found in any language: thus expressions such as “constructions with long-distance dependencies” should not be taken as implying that such patterns necessarily have any mental reality. I will use the terms “template” and “schema” to refer to speakers’ mental representations of these patterns.
where semantic structure is represented in CAPITALS and phonological structure in italics:

(3) YOU THINK SHE WILL GET SOMETHING: WHAT?
    is to  What do you think she’ll get?
    as    SHE HOPES SHE WILL GET SOMETHING: WHAT?
    is to  ???

To solve this problem, the speaker needs to establish correspondences between the relevant parts of semantic and phonological structure: YOU = SHE, therefore the target expression will have the phonological form corresponding to SHE, namely she, in place of the phonological form corresponding to YOU, namely you, and so on. This requires knowledge about linguistic categories (the speaker must know what can be substituted for what), the internal structure of the source expression, i.e., YOU THINK SHE WILL GET SOMETHING: WHAT?/What do you think she’ll get? (so that he/she knows where to substitute it), and about agreement (the auxiliary needs to agree with the new subject). A listener or reader, of course, would have to use the phonological forms of source and target, plus an understanding of the relationships between their constituent parts, to construct a semantic representation of the target.4

If speakers don’t have a ready-made template for non-prototypical questions and have to extrapolate from existing knowledge in the manner described above, such sentences will require extra effort to produce and understand (which should translate into longer processing times) and should be judged to be less acceptable than more prototypical variants. Both of these predictions can be tested experimentally; the present paper is devoted to testing the second one.

An acceptability judgment experiment, of course, can only provide indirect evidence about sentence processing, and hence is less clearly relevant to the subject of this special issue than, for example, a study which compared reading times or the time taken to respond to a sentence. On

3. For ease of expositions, I have assumed that the source expression for the analogy is an actual expression rather than the template; but this need not be the case.
4. Reliance on analogy and schema use are not as different as it might at first seem. As Langacker points out, applying analogy requires the speaker to apprehend an abstract commonality between the source and target forms; and the abstract commonality, of course, is what would be captured by the schema (Langacker 2000: 60; see also Dąbrowska 2008). Furthermore, repeated use of analogy will result in the “abstract commonality” being entrenched until it becomes a linguistic unit in its own right. The critical difference between the two processes, then, is whether the relevant knowledge is retrieved from memory or created “on the fly”. 
the plus side, an acceptability judgment study is easier to conduct (since it
doesn’t require any special apparatus) and hence it is a sensible first step
when investigating a syntactic construction. Furthermore, many linguists
would argue that it provides more useful evidence about the nature of
speakers’ underlying linguistic representations, or “competence” (cf.
Wasow and Arnold 2005), and hence, perhaps, will be less likely to be
dismissed as “mere performance”.

Of course, judging the acceptability of a sentence is a type of perfor-
mance, and, like other types of performance, can be influenced by a vari-
ety of factors: plausibility, complexity, fatigue, mode of presentation, and
so on. This raises an obvious problem for an analyst trying to interpret
the results. The solution to the problem, however, is not—as some lin-
guists have suggested—to give up attempts to study speakers’ judgments
experimentally, but to control as many confounding factors as possible,
and be cautious in interpreting the results.

2. Experimental design

Dąbrowska (2004) reports on a preliminary study showing that speakers
rate prototypical LDD questions such as Where do you think they sent the
documents? and What did you say the burglars stole? as more acceptable
than questions which had lexical subjects, a main verb other than think
or say, and an auxiliary other than do (e.g., Where will the customers
remember they sent the documents? What have the police revealed the
burglars stole?). There was no corresponding effect for declaratives. It is
not clear, however, whether the difference in speakers’ judgments was
due to the choice of subject, verb, or auxiliary, or some combination
of these factors. The experiment described in this paper was designed to
investigate how each of these three factors individually contributes to
speakers’ judgment. It will also examine two additional grammatical fac-
tors: the presence or absence of a complementizer and the number of
clauses intervening between the WH word and the gap, as well as the
effects of plausibility and syntactic complexity.

In the experiment, native speakers of English completed a written ques-
tionnaire in which they were asked to rate the acceptability of LDD ques-
tions of varying degrees of prototypicality. There were seven experimental
conditions:

1. Prototypical LDD questions (WH Prototypical): These had the form
   WH do you think S-GAP? or WH did you say S-GAP?;
2. LDD questions with lexical matrix subjects (WH Subject);
3. LDD questions with auxiliaries other than do (WH Auxiliary);
4. LDD questions with matrix verbs other than think or say (WH Verb);
5. LDD questions with overt complementizers (WH Complementizer);
6. LDD questions with “very long” dependencies, i.e., with an additional complement clause (WH Long);
7. Unprototypical LDD questions (WH Unprototypical): These had a lexical subject, an auxiliary other than do, and a main verb other than think or say, an overt complementizer, and an additional complement clause.

The questionnaire also contained two types of control sentences. Grammatical controls were declarative versions of the LDD questions constructed by replacing the WH word with a noun phrase or a prepositional phrase (and adding a conjunction: see below). Ungrammatical controls involved four types of structures: that trace violations (*That), sentences involving a dependency reaching into a complex NP (*ComplexNP), negative sentences without do support (*Not), and negative sentences with double tense marking (*DoubleTn). Examples of each type of sentence are provided in Table 1; a complete list of all sentences used in one version of the questionnaire is given in the Appendix.

3. Predictions

3.1. General rules

If speakers have the competence attributed to them by generative linguists, and if their grammaticality judgments are a more or less direct reflection of this competence, then we could expect the following prediction to hold:

Prediction 1: Grammatical sentences should receive ratings close to 5; ungrammatical sentences should be rated about 1.

3.2. General rules + processing and pragmatics

Prediction 1 is unrealistic, since it is well known that speakers’ judgments are influenced by factors such as complexity and plausibility, just like any other kind of performance. In particular, sentences involving filler-gap dependencies are computationally more demanding since the filler must be held in memory while the rest of the sentence is being processed (cf. Frazier and Clifton 1989; Hawkins 1999; Kluender and Kutas 1993); and sentences involving a dependency over more than one clause are particularly difficult. Furthermore, since it is rather odd to assert what the addressee thinks or says (and perfectly natural to ask about these things), we might expect an interaction between construction type and the lexical
properties of the matrix subject: specifically, speakers may assign low ratings to declaratives with second person subjects, but accept the corresponding interrogatives.

Taking processing demands and pragmatics into consideration, one might make the following predictions:

**Prediction 2:** WH questions (WH) will receive lower ratings than the corresponding declaratives (DE):

- DE Prototypical > WH Prototypical
- DE Subject > WH Subject
DE Verb > WH Verb
DE Auxiliary > WH Auxiliary
DE Complementizer > WH Complementizer
DE Long > WH Long
DE Unprototypical > WH Unprototypical

Prediction 3: WH questions involving very long dependencies (WH Long, WH Unprototypical) should receive lower ratings than WH questions involving dependencies over just one clause boundary (all other WH conditions). The corresponding declaratives should be rated equally acceptable, since they do not involve long distance dependencies.

WH Prototypical, WH Subject, WH Verb, WH Auxiliary, WH Complementizer > WH Long, WH Unprototypical

Prediction 4: Declaratives with lexical subjects and complement-taking verbs will receive higher ratings than declaratives with second person subjects. There will be no corresponding effect for interrogatives.

DE Subject > DE Prototypical

3.3. Usage-based models

According to the usage-based hypothesis, speakers store lexically specific templates corresponding to frequent combinations they have encountered in their experience such as WH do you think S-GAP? and WH did you say S-GAP? If this hypothesis is correct, prototypical LDD questions (i.e., those that fit one of these templates) should be rated as more acceptable than non-prototypical questions. There should be no corresponding differences in the acceptability of declaratives, or the relevant differences should be smaller.

It was also hypothesised that speakers construct non-prototypical LDD questions on analogy with prototypical questions, by adding elements or substituting a different item in a particular position in a template. Since some elements may be more easily substitutable than others, different substitutions may result in different degrees of acceptability. We know from language acquisition research that children learn to substitute new items into nominal slots relatively early; the ability to substitute verbs into slots emerges later, and auxiliary substitutions are later still (Dąbrowska and Lieven 2005; Lieven et al. 2003, Tomasello et al. 1997). The most likely explanation for this finding is that nominals are autonomous units which can be defined independently of the constructions they occur
in, while verbs are non-autonomous in the sense that their descriptions must make reference to the entities participating in the relationship they designate—in other words, these entities are part of the verb’s profile (cf. Langacker 1987: 298ff). Likewise, tensed auxiliaries, as grounding predications (cf. Langacker 1991: 193ff), conceptually presuppose the events that they ground—designated by the verb plus its arguments. It follows that nominals should be more easily substitutable than verbs, which in turn should be more easily substitutable than auxiliaries.

**Prediction 5**: Prototypical LDD questions will receive higher ratings than questions with lexical subjects, which will be more acceptable than questions with matrix verbs other than *think* or *say*; questions with auxiliaries other than *do* will be judged least acceptable.

WH Prototypical > WH Subject > WH Verb > WH Auxiliary

Finally, LDD questions containing overt complementizers and “very long” dependencies are also less prototypical, in that both contain an extra element (the complementizer or an additional clause). It is not clear whether inserting an extra element is more or less difficult than substitution; however, we can make the following predictions:

**Prediction 6**: LDD questions without overt complementizers will receive higher ratings than questions with *that*:

WH Prototypical > WH Complementizer

**Prediction 7**: LDD questions with very long dependencies (i.e., dependencies reaching over more than one intervening clause) will be judged less acceptable than questions with shorter dependencies, but more acceptable than unprototypical questions (since the latter contain a very long dependency as well as lexical substitutions).

WH Prototypical > WH Long > WH Unprototypical

4. **Method**

4.1. **Stimuli**

4.1.1. *Experimental sentences*. The experimental sentences were constructed by combining a “sentence stub” with a completion consisting of either a complement clause and an adverbial clause or two complement
clauses (see below). The stubs for the WH Prototypical condition were as follows:

(4)  a. What do you think . . .  
     b. Where do you think . . .  
     c. What did you say . . .  
     d. Where did you say . . .

The stubs for the non-prototypical sentences were constructed by changing or adding the relevant element. Thus, in the WH Subject condition, *you* was changed to a proper name (and a third person ending was added to the auxiliary so that it agreed with the subject); in the WH Auxiliary condition, *do* was changed to *will* or *would*; in the WH Verb condition, *think* was replaced with *believe* or *suspect* and *say* with *claim* or *swear*; in the WH Complementizer condition, an overt complementizer (*that*) was added after the verb; and in the WH Unprototypical condition, all of the above changes were made. 5

The completions for conditions 1–5 consisted of a four word complement clause followed by a four-word adverbial clause, e.g.,

(5) (What do you think) the witness will say __
    (stub) (first complement clause)
    if they don’t intervene?
    (adverbial clause)

The completions for sentences with very long dependencies (i.e., conditions 6 and 7) consisted of a two-word complement clause followed by a pronominal subject, verb, and a four-word prepositional phrase, e.g.,

(6) (What do you think) Jo believes
    (stub) (first complement clause)
    he said __ at the court hearing?
    (second complement clause)

Thus, all experimental sentences without complementizers were 12 words long and contained three clauses, with seven words intervening between the WH word and the gap. Sentences with complementizers were 13 words long, with 8 words between the WH word and the gap.

There were two versions of the questionnaire, and two sets of completions. In version 1, the stubs were combined with completion set 1 in con-

5. Throughout this paper, I use the term *non-prototypical* to refer to questions which differ in some respect from the prototypical instances of the construction, and the term *unprototypical* to refer to questions which differ from the prototype in all relevant respects.
ditions 1, 3, 5 and 7 and with completion set 2 in conditions 2, 4, and 6; in version 2, the stubs were combined with completion set 1 in even-numbered conditions and with completion set 2 in odd-numbered conditions.

4.1.2. Grammatical controls. The grammatical control conditions were constructed by supplying appropriate lexical material for the WH word in the appropriate position in the clause; in addition, a conjunction (so, and, or but) was added at the beginning of declaratives corresponding to interrogatives with the auxiliary do: for example, the declarative counterpart of

(7) What do you think the witness will say if they don’t intervene?

was

(8) But you think the witness will say something if they don’t intervene.

This was done so that the interrogative sentence and the corresponding declarative control contained the same number of words; it also made the declarative sentences sound more natural. (It is somewhat odd for a speaker to assert what the addressee thinks or said; adding the conjunction makes the sentence pragmatically more plausible because it conveys the impression that the speaker is either inferring the addressee’s beliefs from his or her words, contrasting them with those of another person, or clearing up an apparent misunderstanding).

4.1.3. Ungrammatical controls. The ungrammatical control conditions, like the experimental sentences and the grammatical controls, contained subordinate clauses. Half of the ungrammatical sentences were declarative and the other half interrogative. That-trace sentences (*That) contained an overt complementizer immediately before a gap in the subject position:

(9) *What do you think that __ probably got lost during the move?
(10) *Who do you think that __ will turn up in the evening?

In Complex NP sentences (*ComplexNP) the matrix clause contained a complement-taking noun (claim, fact, rumour, hypothesis) followed by a complement clause with a gap, e.g.,

(11) *What did Claire make the claim that she read __ in a book?
(12) *Where did you discover the fact that the criminals put the car __?

In Negatives without do support (*Not), the matrix clause contained a negated verb but no auxiliary, with tense being marked on the main verb:
Her husband not claimed they asked where we were going.

Finally, in declaratives with double tense/agreement marking (*DoubleTn), the matrix clause contained a third person subject and a negated verb, and agreement was marked on the auxiliary as well as on the main verb:

*The girl doesn’t remembers where she spent her summer holidays.

The same ungrammatical controls were used in both versions of the questionnaire. There were four sentences in each condition, giving a total of 16 ungrammatical controls.

4.1.4. Constructing the questionnaire. The test sentences were divided into four blocks, each containing one sentence from each of the seven experimental and eleven control conditions. The order of the sentences within each block was random.

A full list of the sentences used in version 1 of the experiment is given in the Appendix.

4.2. Participants

38 second and third year literature students from the School of English at the University of Newcastle participated in the experiment. All were native speakers of English.

4.2.1. Procedure. Participants were asked to complete a written questionnaire and were given the following instructions:

The questionnaire is part of a study of speakers’ intuitions about English sentences. It is not an intelligence test or a grammar test.

Please indicate how acceptable/unacceptable you find each of the following sentences by choosing a number on a scale from 1 (very bad) to 5 (fine). Read the sentences carefully, but do not spend too much time thinking about them: we are interested in your initial reaction. Do not go back and change your responses to earlier sentences.

The instructions were followed by two examples for which ratings were provided:

(15) Will the girl who won the prize come to the party?
(16) Did the man who arrive by train is my cousin?

The first was given a rating of 5 and the second 1. These examples were provided in order to anchor the participants’ ratings. Thus, in essence, the
participants’ task was to decide whether the sentences in the questionnaire were more like 1, more like 5, or in-between.

The questionnaires were distributed after a lecture and took 10–15 minutes to complete. Participants were randomly assigned to one version of the questionnaire, with about half completing each version.

5. Results and discussion

5.1. Grammatical v. ungrammatical sentences

The mean acceptability ratings for all conditions are given in Table 2. For clarity, the same information is presented visually in Figure 1, where the bars corresponding to each experimental condition have been arranged from highest to lowest. Although the mean rating for all grammatical sentences combined (3.67) was considerably higher than for the ungrammatical controls (1.98), there is no sharp contrast between grammatical and ungrammatical sentences. What we have instead is a continuum of acceptability ratings ranging from 4.3 for prototypical WH questions to 1.3 for negatives without do support, with the other sentence types occupying various intermediate points. The four ungrammatical sentence types cluster at the lower end of the continuum; however, the acceptability ratings for unprototypical LDD questions were not

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.59</td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.71</td>
</tr>
<tr>
<td>WH Auxiliary</td>
<td>3.23</td>
<td>0.83</td>
</tr>
<tr>
<td>WH Complementizer</td>
<td>3.84</td>
<td>0.84</td>
</tr>
<tr>
<td>WH Long</td>
<td>3.85</td>
<td>0.76</td>
</tr>
<tr>
<td>WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
</tr>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
</tr>
<tr>
<td>DE Subject</td>
<td>4.00</td>
<td>0.63</td>
</tr>
<tr>
<td>DE Verb</td>
<td>3.74</td>
<td>0.78</td>
</tr>
<tr>
<td>DE Auxiliary</td>
<td>3.49</td>
<td>0.66</td>
</tr>
<tr>
<td>DE Complementizer</td>
<td>3.53</td>
<td>0.79</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
</tr>
<tr>
<td>*That</td>
<td>2.50</td>
<td>0.75</td>
</tr>
<tr>
<td>*DoubleTn</td>
<td>2.41</td>
<td>0.95</td>
</tr>
<tr>
<td>*ComplexNP</td>
<td>1.69</td>
<td>0.56</td>
</tr>
<tr>
<td>*Not</td>
<td>1.31</td>
<td>0.49</td>
</tr>
</tbody>
</table>
significantly different from those for that-trace and double-tense sentences (WH Unprototypical v. *that: $t(37) = 0.26, p = 0.798$; WH Unprototypical v. *DoubleTn: $t(37) = 0.70, p = 0.486$), although they were higher than those for the other two ungrammatical control conditions (WH Unprototypical v. *Complex NP: $t(37) = 6.15, p < 0.001$; WH Unprototypical v. *Not: $t(37) = 8.81, p < 0.001$). Thus, the pure competence grammar prediction that grammatical sentences will receive ratings close to 5 and ungrammatical sentences will be rated about 1 is clearly false.

5.2. Grammatical sentences

A preliminary analysis of the participants’ ratings for the grammatical sentences was conducted using a construction (2) × prototypicality (7) × version (2) ANOVA. The analysis revealed a significant main effect of prototypicality, $F(6, 216) = 49.82, p < 0.001$, $\eta^2_p = 0.58$, and a prototypicality × construction interaction, $F(6, 216) = 15.14, p < 0.001$, $\eta^2_p = 0.30$. No other effect or interaction was significant. Since there was no significant effect of version, and no interaction between version and any of the other factors, the results for the two versions were collapsed in all further analyses.
5.2.1. **Processing cost of dependencies: Questions v. declaratives.** Prediction 2 was that LDD questions will receive lower ratings than the corresponding declaratives because they involve displaced constituents. This prediction was not confirmed: there was no significant effect of construction. Instead, as indicated above and shown in Table 3, we have an interaction between construction type and lexical content, with prototypical LDD questions, questions with lexical subjects, and questions with overt complementizers being judged significantly more acceptable than the corresponding declaratives. (Note that the significance levels reported in Table 3 and elsewhere in this paper have not been corrected for multiple comparisons: since the hypothesis tested predicts that all the relevant comparisons should be significant, using the Bonferroni adjustment or an equivalent method would not be appropriate.)

The interrogative sentences, particularly in the WH Prototypical condition, contained some frequent bigrams (*what do, do you, you think*): these occur with a frequency of 6433, 27602, and 9901 respectively in the British National Corpus. This could be partly responsible for the fact that interrogatives were rated as more acceptable than the corresponding declaratives in most conditions. However, as shown in Table 4, there is no strong relationship between the number of frequent bigrams and acceptability ratings of WH questions (i.e., questions containing more high-frequency bigrams are not necessarily more acceptable) or the num-

---

**Table 3. Testing prediction 2**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Mean</th>
<th>SD</th>
<th>t-test value</th>
<th>p value</th>
<th>Prediction confirmed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>5.234</td>
<td>&lt;0.001</td>
<td>✖✖</td>
</tr>
<tr>
<td>&lt;DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.59</td>
<td>2.859</td>
<td>0.007</td>
<td>✖✖</td>
</tr>
<tr>
<td>&lt;DE Subject</td>
<td>4.00</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Complementizer</td>
<td>3.84</td>
<td>0.84</td>
<td>2.149</td>
<td>0.038</td>
<td>✖✖</td>
</tr>
<tr>
<td>&lt;DE Complementizer</td>
<td>3.53</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.71</td>
<td>1.644</td>
<td>0.109</td>
<td>✖</td>
</tr>
<tr>
<td>&lt;DE Verb</td>
<td>3.74</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td>0.452</td>
<td>0.654</td>
<td>✖</td>
</tr>
<tr>
<td>&lt;DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Auxiliary</td>
<td>3.23</td>
<td>0.83</td>
<td>1.950</td>
<td>0.059</td>
<td>✖</td>
</tr>
<tr>
<td>&lt;DE Auxiliary</td>
<td>3.49</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td>5.882</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&lt;DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ✓ indicates that a prediction has been confirmed; ✖ indicates that a prediction has not been confirmed; ✖✖ indicates a significant difference in the opposite direction.
ber of frequent bigrams and the advantage for questions over declaratives
(questions containing more high-frequency bigrams are not necessarily
better than the corresponding declaratives). Furthermore, bigram fre-
quency cannot explain the interaction between construction type and
verb or between construction type and complementizer (see below), since
the words immediately preceding and following the verb and the com-
plementizer were the same in both the declarative and the interrogative var-
ants. In fact, for sentences with complementizers, bigram frequency
makes precisely the wrong predictions. In the WH Prototypical and DE
Prototypical conditions, the main clause verb (think or say) was followed
by the pronoun they or we or the determiner the, while in the WH Com-
plementizer and DE Complementizer conditions, it was followed by the
complementizer that. The mean frequency of the bigrams think the, think
they, think we, say the, say they, say we in the British National Corpus is
2539, while the mean frequency of the bigrams think that and say that is
9473. Thus, if acceptability ratings were simply a reflection of bigram fre-
quency, sentences with complementizers should receive higher ratings
than sentences without them. (The mean frequency of the bigrams that
they, that the and that we is even higher: 59333.) However, as we will see
in section 5.2.4, the ratings for WH questions with complementizers were

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean</th>
<th>Difference score*</th>
<th>Frequent bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.74</td>
<td>what do you think</td>
</tr>
<tr>
<td>WH Complementizer</td>
<td>3.84</td>
<td>0.31</td>
<td>what do you think</td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.25</td>
<td>what do you think</td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.19</td>
<td>what do you</td>
</tr>
<tr>
<td>WH Long</td>
<td>3.85</td>
<td>−0.04</td>
<td>what do you think</td>
</tr>
<tr>
<td>WH Auxiliary</td>
<td>3.23</td>
<td>−0.26</td>
<td>you think</td>
</tr>
<tr>
<td>WH Unprototypical</td>
<td>2.54</td>
<td>−0.60</td>
<td></td>
</tr>
</tbody>
</table>

* Difference scores were computed by subtracting the rating of the declarative control from
the rating of the interrogative sentence.
considerably lower than for the prototypical variants, while there was no difference in the acceptability of the corresponding declaratives.

5.2.2. **Processing cost of “very long” dependencies.** According to prediction 3, WH questions containing “very long” dependencies, i.e., dependencies spanning two clause boundaries, should receive lower ratings than questions containing dependencies across just one clause boundary, whilst the corresponding declaratives should be rated equally acceptable, since they do not involve a filler-gap dependency. This prediction was evaluated by comparing the mean ratings for questions with “very long” dependencies (WH Long and WH Unprototypical) and the corresponding declaratives and for questions containing dependencies over one clause boundary (WH Prototypical, WH Subject, WH Verb, WH Auxiliary, WH Complementizer) and the corresponding declaratives. The ratings were analyzed using a 2 × 2 ANOVA with the within-participants factors of construction type (WH question, declarative) and complementation (1 clause, 2 clauses). This revealed a main effect of complementation, \( F(1,37) = 50.44, \ p < 0.001, \ \eta_p^2 = 0.58 \), indicating that sentences containing two complement clauses were judged as less acceptable than sentences containing one complement clause and one adverbial clause. This was qualified by a construction × complementation interaction, \( F(1,37) = 34.14, \ p < 0.001, \ \eta_p^2 = 0.48 \) (see Figure 2). Post-hoc comparisons showed that WH questions with very long dependencies were judged significantly worse than questions with long dependencies: \( t(37) = 9.62, \ p < 0.001 \).

![Figure 2. Construction type × complementation interaction (All conditions)](image-url)
For declaratives, the corresponding difference approaches significance: $t(37) = 1.87$, $p = 0.070$. This suggests that processing difficulty had the predicted effect on acceptability ratings.

However, comparisons of ratings for each of two “very long” dependency questions with each of the “long” dependency conditions suggests a slightly more complex picture. As shown in Table 5, WH Long items were judged significantly worse than WH Prototypical and WH Subject items, and unprototypical questions were judged significantly worse than all other WH questions. However, there was no significant difference between WH Long and WH Verb items or between WH Long and WH Complementizer (indeed, the latter received slightly higher ratings), and questions with a modal auxiliary were judged significantly worse than WH Long sentences. Since the declarative versions of unprototypical LDD questions were also judged to be less acceptable than the other declarative sentences (see Table 6), the low acceptability ratings for the unprototypical sentences may be partially due to the lexical content of the sentences. Thus, although the existence of very long dependencies may have an effect on acceptability, this can only account for some of the observed differences.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Mean</th>
<th>SD</th>
<th>t-test value</th>
<th>p value</th>
<th>Prediction confirmed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>4.67</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.59</td>
<td>3.63</td>
<td>0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.71</td>
<td>0.62</td>
<td>0.539</td>
<td>x</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Auxiliary</td>
<td>3.23</td>
<td>0.83</td>
<td>4.13</td>
<td>&lt;0.001</td>
<td>xx</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Complementizer</td>
<td>3.84</td>
<td>0.84</td>
<td>0.11</td>
<td>0.910</td>
<td>x</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>14.17</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.59</td>
<td>15.36</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.71</td>
<td>11.29</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Auxiliary</td>
<td>3.23</td>
<td>0.83</td>
<td>5.73</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Complementizer</td>
<td>3.84</td>
<td>0.84</td>
<td>9.38</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2.3. Pragmatics. Prediction 4 stated there should be an interaction between construction type and the lexical properties of the matrix subject: declaratives with lexical subjects should receive higher ratings than declaratives with second person subjects, but there should be no corresponding difference for questions. To test this prediction, a $2 \times 2$ ANOVA with the within-participants factors of construction (WH question, declarative) and subject (second person, lexical). The analysis showed that the predicted interaction did indeed occur: $F(1, 37) = 14.82$, $p < 0.001$, $\eta_p^2 = 0.29$ (see Figure 3); the main effect of construction was also significant, $F(1, 37) = 25.01$, $p < 0.001$, $\eta_p^2 = 0.40$. Further analysis confirmed that declaratives with lexical subjects were judged more acceptable than declaratives with second person subjects: $t(37) = 3.92$, $p < 0.001$. The ratings for interrogatives with lexical subjects were marginally lower than for interrogatives with second person subject, but the difference was not statistically significant: $t(37) = 0.91$, $p = 0.368$.

5.2.4. Prototypicality effects. According to the usage-based hypothesis, any modification of the LDD template should result in lower

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Mean</th>
<th>SD</th>
<th>$t$-test value</th>
<th>$p$ value</th>
<th>Same as question?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td>3.10</td>
<td>0.004</td>
<td>x</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Subject</td>
<td>4.00</td>
<td>0.63</td>
<td>1.22</td>
<td>0.230</td>
<td>x</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Verb</td>
<td>3.74</td>
<td>0.78</td>
<td>1.45</td>
<td>0.156</td>
<td>✓</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Auxiliary</td>
<td>3.49</td>
<td>0.66</td>
<td>3.89</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Complementizer</td>
<td>3.53</td>
<td>0.79</td>
<td>3.68</td>
<td>0.001</td>
<td>x</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td>3.22</td>
<td>0.003</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Subject</td>
<td>4.00</td>
<td>0.63</td>
<td>6.74</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Verb</td>
<td>3.74</td>
<td>0.78</td>
<td>4.31</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Auxiliary</td>
<td>3.49</td>
<td>0.66</td>
<td>2.99</td>
<td>0.005</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Complementizer</td>
<td>3.53</td>
<td>0.79</td>
<td>3.17</td>
<td>0.003</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
acceptability ratings for interrogative sentences but have no effect (or a much smaller effect) on declaratives. In other words, usage-based theories predict an interaction between construction type and lexical content.

As we have just seen, although there is a significant interaction between construction type and the lexical properties of the subject, this is best interpreted as reflecting pragmatic implausibility of second person declaratives with think and say. Substituting a lexical subject for you in questions did not result in a significant reduction of acceptability, although there was a small difference in the predicted direction. This suggests that the LDD template does not specify the subject—although it is also possible that the processing cost of NP substitution is too small to be revealed by an acceptability judgment task.

The relationship between construction type and the other four factors (verb, auxiliary, complementizer, and the number of complement clauses) was investigated by means of four additional 2 × 2 ANOVAs (see Table 7) followed up by t-tests. All the interactions were as predicted by the usage-based account. There was a significant interaction between construction type and verb (see Figure 4): changing the matrix verb in a LDD question results in significantly lower acceptability ratings (t(37) = 3.23, p = 0.003); changing the verb in the corresponding declarative, on the other hand, results in a slightly more acceptable sentence, although the difference is not statistically significant (t(37) = 1.45, p = 0.155). There was also an interaction between construction type and auxiliary (see Figure 5): replacing do or did with the modal auxiliary will
or *would* made the interrogatives less acceptable \((t(37) = 8.26, p < 0.001)\), while adding the same auxiliaries in declaratives had no effect on ratings \((t(37) = 0.56, p = 0.578)\). The size of the interaction between construction type and complementizer (Figure 6) was somewhat smaller, although it was also in the predicted direction: adding an overt complementizer had no effect on declaratives \((t(37) = 0.44, p = 0.666)\) but
resulted in lower ratings for interrogatives ($t(37) = 3.77$, $p = 0.001$). Finally, there is a significant interaction between construction type and complementation: adding a second complement clause between the filler and the gap makes the interrogative less acceptable while the corresponding declarative is more acceptable (see Figure 7; for pairwise comparisons, see Table 5). (Note that this analysis compares just the WH Prototypical and WH Long conditions and their declarative counterparts, i.e., sentences which are most closely matched lexically. The analysis in section 5.2.2 contrasted questions with “very long” dependencies, i.e., WH Long and WH Unprototypical, with questions containing dependencies spanning just one clause, i.e., WH Prototypical, WH Subject, WH Verb, WH Auxiliary, and WH Complementizer.)

The usage-based model adopted here also predicts a particular order in the acceptability of non-prototypical LDD questions: specifically, LDD questions with a lexical subject should be more acceptable than questions with a non-prototypical matrix verb, which in turn should be more

---

6. We know from the psycholinguistic literature that overt complementizers facilitate processing, presumably because they signal the presence of a subordinate clause and thus help the processor to avoid garden path effects: for instance, sentences with complement clauses introduced by a complementizer are read faster than sentences without complementizers, even when the main clause verb has a strong preference for clausal complements (Trueswell et al. 1993, Holmes et al. 1989). Thus the presence of the complementizer effect for questions provides strong evidence in favour of lexical storage of the whole construction.
acceptable than those with a non-prototypical auxiliary. As shown in Tables 8 and 9, these predictions have also been confirmed. (Table 8 also shows the results of pairwise comparisons relevant for testing other usage-based predictions for the sake of completeness.)

Although the model made no specific predictions about the relative size of the effect of the other manipulations, it is interesting to see how they compare to lexical substitutions in the subject, verb and auxiliary slot.
As can be seen from Figure 1, the effects of adding a complementizer and of adding an additional complement clause are about the same as that of changing the matrix verb. The mean acceptability ratings for WH Verb, WH Long and WH Complementizer sentences were 3.93, 3.85, and 3.84 respectively; none of the differences between these conditions was statistically significant (WH Verb v. WH Complementizer: $t(37) = 0.71$, $p = 0.484$; WH Verb v. WH Long: $t(37) = 0.62$, $p = 0.539$; WH Complementizer v. WH Long: $t(37) = 0.11$, $p = 0.910$).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Mean</th>
<th>SD</th>
<th>$t$-test value</th>
<th>p value</th>
<th>Prediction confirmed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>0.91</td>
<td>0.368</td>
<td>x</td>
</tr>
<tr>
<td>&gt;WH Subject</td>
<td>4.25</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Subject</td>
<td>4.25</td>
<td>0.59</td>
<td>3.32</td>
<td>0.002</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Verb</td>
<td>3.93</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Verb</td>
<td>3.93</td>
<td>0.71</td>
<td>4.96</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Auxiliary</td>
<td>3.23</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>3.77</td>
<td>0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Complementizer</td>
<td>3.84</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Prototypical</td>
<td>4.31</td>
<td>0.63</td>
<td>4.69</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WH Long</td>
<td>3.85</td>
<td>0.76</td>
<td>10.51</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>&gt;WH Unprototypical</td>
<td>2.54</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Mean</th>
<th>SD</th>
<th>$t$-test value</th>
<th>p value</th>
<th>Same as question?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td>3.92</td>
<td>&lt;0.001</td>
<td>x</td>
</tr>
<tr>
<td>DE Subject</td>
<td>4.00</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Subject</td>
<td>4.00</td>
<td>0.63</td>
<td>2.85</td>
<td>0.007</td>
<td>✓</td>
</tr>
<tr>
<td>DE Verb</td>
<td>3.74</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Verb</td>
<td>3.74</td>
<td>0.78</td>
<td>2.26</td>
<td>0.030</td>
<td>✓</td>
</tr>
<tr>
<td>DE Auxiliary</td>
<td>3.49</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td>0.44</td>
<td>0.666</td>
<td>x</td>
</tr>
<tr>
<td>DE Complementizer</td>
<td>3.53</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Prototypical</td>
<td>3.57</td>
<td>0.85</td>
<td>3.10</td>
<td>0.004</td>
<td>x</td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE Long</td>
<td>3.89</td>
<td>0.75</td>
<td>8.44</td>
<td>&lt;0.001</td>
<td>✓</td>
</tr>
<tr>
<td>DE Unprototypical</td>
<td>3.14</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3. Why are “real life” LDD questions so stereotypical?

Why are questions with long distance dependencies so stereotypical? One possible explanation is offered by Verhagen (2005). Verhagen observes that the propositional content of most complementation constructions is expressed by the subordinate clause; the main clause normally just signals epistemic stance (see also Thompson 2002), i.e., it invites the hearer to adopt a particular subjective perspective on the object of conceptualization. The greater the “distance” between the onstage conceptualizer (i.e., the subject of the main clause) and the ground (in the sense of Langacker 1987), the more difficult it is to construe the main clause as an epistemic marker (as opposed to a prediction in its own right). Verhagen argues that this distance is minimal when the conceptualizer is the first person (in declaratives) or the second person (in interrogatives), when the verb is relatively generic, and when there are no other elements qualifying the verb; it follows that the matrix clause in LDD questions will normally contain a second person subject, a relatively non-specific verb such as think and say, and no additional constituents.

A different, but not necessarily incompatible, explanation for restrictions on questions and other constructions with long distance dependencies is proposed by Goldberg (2006). Goldberg argues that differences in acceptability arising as a result of the use of different main clause verbs can be explained by appealing to a general principle which she calls BCI, which states that backgrounded constituents are islands. The gap in a filler-gap dependency construction must occur within the “potential focus domain”; the constituent containing the gap cannot be backgrounded. Since complements of factive verbs and manner of speaking verbs (as well as complex NPs, sentential subjects, and presupposed adjuncts) are backgrounded, they cannot participate in filler-gap constructions.

An experimental study by Ambridge and Goldberg (this issue) provides some empirical support for this proposal. Participants in this experiment completed two tasks. In the first task, they rated the acceptability of WH questions with long distance dependencies (e.g., What did Jess think that Dan liked?) and the corresponding declaratives (e.g., Daniele thought that Jason liked the cake). In the second task they were presented with a negated sentence containing a verb complement clause (e.g., Maria didn’t know that Ian liked the cake) and asked to judge to what extent it implied the negation of the complement clause (Ian didn’t like the cake): this measured the extent to which speakers judged the information in the subordinate clause to be presupposed, and thus backgrounded. The main finding was that, as predicted by the BCI hypothesis, there was a very strong negative correlation between responses on the negation test and “difference
scores” computed by subtracting the acceptability rating of the questions from those of the corresponding declaratives, and a weaker negative correlation between responses on the negation test and acceptability ratings. It remains to be seen whether BCI can also explain other restrictions on LDD questions documented in this study: the fact that they strongly disprefer main clause verbs other than think or say (not just factives and manner-of-speaking verbs), auxiliaries other than do, and complementizers, and that in real life (as opposed to the examples found in the linguistic literature) they virtually never involve a dependency spanning more than one clause.

Thus we have two independent proposals explaining why particular lexical variants of LDD questions may be preferred or dispreferred in usage. A central claim of usage-based approaches is that mental grammars are shaped by usage patterns: it is thus not surprising that speakers develop strong lexically specific templates for LDD questions and possibly fail to develop more abstract representations of these constructions.

6. Conclusion

The most striking result of the experiment reported here is the existence of strong prototypicality effects for LDD questions. Prototypical instances of this construction, i.e., those which fit one of the templates postulated on the basis of corpus research (WH do you think S-GAP?, WH did you say S-GAP?) were judged to be the most acceptable of all sentences. Departures from the prototype (use of a different auxiliary or verb in the matrix clause, addition of a complementizer or an extra complement clause) resulted in lower acceptability ratings. Crucially, there was no corresponding effect on declaratives, so the differences in grammaticality cannot be attributed simply to the properties of the lexical items used in the experiment. Acceptability also depended on the type of substitution required: nominals are apparently easier to substitute than verbs, which in turn were easier than auxiliaries.

The participants’ judgments were also influenced by pragmatic considerations: declaratives with lexical subjects (DE Subject, e.g., So Steve said the children could stay here when their father returns) were judged to be more acceptable than declaratives with second person subjects (DE Prototypical, e.g., So you said the children could stay here when their father returns). This effect, however, was fairly small in comparison with the purely lexical effects.

Adding an additional complement clause also reduced the acceptability of the questions (and had the opposite effect on declaratives). This could be attributed to the greater processing demands posed by the increased
syntactic distance between the filler and the gap, since the filler must be held in working memory while the pre-gap part of the sentence is being processed. However, questions with very long dependencies (with two clause boundaries intervening between the filler and the gap) were not consistently judged to be less acceptable than questions involving dependencies across only one clause boundary; and the effect of adding an additional complement clause was no bigger than that of adding a complementizer or changing the matrix verb. Thus, appealing to prototypicality effects provides a more parsimonious explanation for these findings. This is not to say that the processing demands of holding the filler in memory have no effect on processing—but the costs may be relatively small compared the effects of prototypicality.\(^7\)

Interestingly, unprototypical LDD questions—those with a complementizer, an additional verb complement clause, a lexical subject, a modal auxiliary, and a verb other than *think* or *say* in the main clause—were judged to be just as bad as *that*-trace violations and sentences with double tense/agreement marking (though better than sentences involving “extraction” out of a complex NP and negatives without an auxiliary). This is consonant with the results of two acceptability experiments conducted by Kluender and Kutas (1993). In their first experiment, in which participants were required to provide speeded categorical acceptability judgments, LDD questions were accepted only 54 percent of the time. In the second experiment, participants rated acceptability on a scale from 1 to 40, and could take as much time as they wished to make the judgment. The mean acceptability rating for LDD questions was 19—significantly higher than that for WH island variations, but much lower than those for Y/N questions containing complement clauses. Kluender and Kutas conclude that the low acceptability of LDD questions is attributable to the processing demands of holding the filler in working memory. However, since their stimuli were quite different from prototypical LDD questions (they all contained overt complementizers and non-prototypical verbs; some also had lexical subjects or auxiliaries other than *do*), it could also be explained by appealing to their unprototypicality.

So what does the presence of prototypicality effects in acceptability judgments about LDD questions—in particular, those due to the lexical

---

7. Acceptability judgment and ease of processing are of course two different things, although they tend to be correlated: other things being equal, sentences which are difficult to process tend to be judged less acceptable (cf. Fanselow and Frisch 2006; Frazier and Clifton 1989; Kluender and Kutas 1993). Note, too, that unprototypical LDD questions contain more dysfluencies than prototypical instances of the construction (Dąbrowska forthcoming), which also suggests that they are more difficult to process.
properties of the sentences, since these are easier to interpret—tell us about speakers’ mental representations of this construction? It has long been acknowledged, of course, that the choice of lexical items in a sentence affects speakers’ acceptability judgments. In the generative tradition, this is usually regarded as a confound: the presence of a particular word can make an otherwise well-formed sentence unacceptable, which could lead the analyst to draw incorrect conclusions about grammar; lexical effects, therefore, are something that should be “controlled for where possible, discounted when encountered” (Featherston 2005: 702). In this case, however, we are dealing with the opposite situation: LDD questions are fully acceptable only with particular lexical content. This suggests that they are more like a constructional idiom than a fully general construction. In other words, questions with long-distance dependencies are conventional units with an unusual form (a WH word at the beginning of the main clause associated with a gap in a subordinate clause) and a specialized meaning: the unit \textit{WH do you think S-GAP?} is used to inquire about the speakers’ opinion about the content of the subordinate clause (\textit{What do you think he wants} \approx “In your opinion, what does he want?”); and \textit{WH did you say S-GAP?} is used when addressee already gave the speaker the relevant information but the speaker does not remember (\textit{When did you say he came?} \approx “You’ve already told me when he came, but please tell me again”.)\(^8\) The non-prototypical uses are rather like what Moon (1998) calls “exploitations” of idioms exemplified by expressions such as \textit{throw in the moist towelette} (constructed on analogy with \textit{throw in the towel}) or \textit{use an earthmover to crack a nut} (cf. \textit{use a sledgehammer to crack a nut}). Once all the lexical content has been changed (cf. the WH Unprototypical condition), it is no longer possible to identify the motivating construction, and hence the sentence is judged as unacceptably low.\(^9\) Alternatively, one could argue that when processing unprototypical LDD questions, speakers have to fall back on more abstract schemas (the mental analogues of the WH question construction and the complementation construction), and that the low acceptability ratings for

---

\(^8\) It is interesting to note that the CollinsCobuild English Language Dictionary (Sinclair 1987: 1519) lists the use of \textit{think} in long-distance questions as a separate sense of the verb.

\(^9\) With most idioms, substituting different lexical items for every word would result in expressions which are still judged acceptable, provided they make sense semantically: for example, taking \textit{use a sledgehammer to crack a nut} as the model, one could construct perfectly acceptable phrases such as \textit{do a headstand to impress a neighbour} and \textit{tickle an earthworm to amuse the children}. This is possible because speakers have fully general schemas for the transitive and the infinitival construction.
such sentences reflect the difficulty of combining highly complex and abstract schemas.

At this point one may wonder why the possibility that speakers have lexically specific knowledge about LDD questions has not even been considered by most syntacticians in spite of the fact that these constructions have been so intensively studied for several decades. The short answer is that the possible existence of lexical effects was not regarded as theoretically interesting—so although there are a number of experimental studies of LDD constructions (see, for example, Cowart 1997; Frazier and Clifton 1989; Kluender and Kutas 1993), to my knowledge, nobody has systematically investigated the effect of lexical content on speakers’ linguistic intuitions about them. A second reason is that linguists tend to rely on their own intuitions—and linguists’ intuitions about LDD questions may be systematically different from those of ordinary speakers. Da˛browska (in press b) shows that linguists tend to judge unprototypical LDD questions as considerably more acceptable than that trace violations and sentences with double tense marking, and not much worse than prototypical questions. This could be a reflection of their theoretical commitments (the belief that instances of “the same” construction should be equally grammatical), but it could also be a result of differences in linguistic experience. Many linguists spend a considerable amount of time constructing examples of the structures they are interested in and reading papers containing such constructed examples. Since LDD questions have been the object of very intensive research, it is likely that linguists (or at least linguists who work on LDD constructions, or discuss them with their students) have been exposed to more instances of this construction than most ordinary language users, and, crucially, the instances they have encountered are much more varied, as demonstrated by the examples in (1).

10. There is some work on lexical effects on acceptability judgments in basic argument structure constructions: see Theakston 2004, Ambridge et al. in press. In both studies, speakers judged argument structure violations with high frequency verbs (e.g., *I poured you with water) as less acceptable than argument structure violations with low frequency verbs (e.g., *I dribbled teddy with water); the authors explain this by appealing to the higher entrenchment of the pattern with the frequent verb. Ambridge et al. (2008) also found that fully grammatical sentences with high frequency verbs were judged slightly more acceptable than sentences with low-frequency verbs, although the difference was very small (for adults, 4.82 v. 4.76; the authors do not indicate whether or not it was statistically significant).

11. Note that constructing examples for a linguistic paper (or for one’s students to analyse) is a very different kind of activity from ordinary language use: it is conscious and deliberate, and relies on metalinguistic and/or general problem-solving abilities rather than normal linguistic routines.
As a result, they are much more likely to develop more general representations of these constructions, and accept unprototypical instances of them.12

Received 11 April 2007 Sheffield University, UK
Revision received 18 December 2007

Appendix: List of sentences used in version 1 of the experiment

Experimental sentences

Prototypical LDD question (WH Prototypical)
  What did you say the family should know before they go there?
  What do you think they decided to do when they got home?
  Where did you say they hid the treasure when they found out?
  Where do you think the children could stay when their father returns?

LDD question with lexical subject (WH Subject)
  What did Steve say we bought for Alice when we visited her?
  What does Claire think the witness will say if they don’t intervene?
  Where did Andy say the young man went after they found her?
  Where does Paul think we put the documents after he saw them?

LDD question with a different verb (WH Verb)
  What did you claim we bought for Alice when we visited her?
  What do you believe the witness will say if they don’t intervene?
  Where did you swear the young man went after they found her?
  Where do you suspect we put the documents after he saw them?

LDD question with a different auxiliary (WH Auxiliary)
  What will you say the family should know before they go there?
  What would you think they decided to do when they got home?
  Where will you say they hid the treasure when they found out?
  Where would you think the children could stay when their father returns?

LDD question with an overt complementizer (WH Complementizer)
  What did you say that the family should know before they go there?
  What do you think that they decided to do when they got home?
  Where did you say that they hid the treasure when they found out?

12. For other work suggesting that linguists’ judgments may be systematically different from those of ordinary speakers, see Spencer 1973 and Bradac et al. 1980. See also Hiramatsu 1999 and Snyder 2000 for experimental studies of “syntactic satiation”, a phenomenon whereby sentences which were initially judged ungrammatical become increasingly acceptable as a result of repeated exposure.
Where do you think that the children could stay when their father returns?

**LDD question with an additional subordinate clause (WH Long)**
What did you say Eve claimed we bought during our first visit?
What do you think Jo believes he said at the court hearing?
Where did you say Mike swore he went after the evening performance?
Where do you think Phil suspects we were during the last war?

**Unprototypical LDD question (WH Unprototypical)**
What will Steve believe that Jo thinks they did with their old furniture?
What would Claire claim that Eve said they know about the whole affair?
Where will Andy suspect that Phil thinks they stayed during the school holidays?
Where would Paul swear that Mike said they were during the afternoon session?

**Grammatical control sentences**

***Prototypical*** declarative (DE Prototypical)
And you think the children could stay here when their father returns.
But you think they decided to do something when they got home.
So you said the family should know everything before they go there.
So you said they hid the treasure somewhere when they found out.

Declarative with lexical subject (DE Subject)
And Claire thinks the witness will say something if they don’t intervene.
But Steve said we bought something for Alice when we visited her.
So Andy said the young man went home after they found her.
So Paul thinks we put the documents back after he saw them.

Declarative with a different verb (DE Verb)
And you swore the young man went home after they found her.
But you suspect we put the documents back after he saw them.
So you believe the witness will say something if they don’t intervene.
So you claimed we bought something for Alice when we visited her.

Declarative with a different auxiliary (DE Auxiliary)
You will say the family should know everything before they go there.
You will say they hid the treasure somewhere when they found out.
You would think the children could stay here when their father returns.
You would think they decided to do something when they got home.

Declarative with an overt complementizer (DE Complementizer)
And you said that the family should know everything before they go there.
But you think that the children could stay here when their father returns. So you said that they hid the treasure somewhere when they found out. So you think that they decided to do something when they got home.

*Declarative with an additional subordinate clause (DE Long)*
And you think Phil suspects we were here during the last war. But you said Mike swore he went home after the evening performance. So you said Eve claimed we bought something during our first visit. So you think Jo believes he said something at the court hearing.

*‘Unprototypical’ Declarative (DE Unprototypical)*
Andy will suspect that Phil thinks they stayed here during the school holidays. Claire would claim that Eve said they know everything about the whole affair. Paul would swear that Mike said they were here during the afternoon session. Steve will believe that Jo thinks they did something with their old furniture.

*Ungrammatical control sentences that trace sentences (*that)*
What did you say that will kill cockroaches but not ants?
What do you think that probably got lost during the move?
Who did you say that ate the spinach your mother cooked?
Who do you think that will turn up in the evening?

*Sentences with extraction from a complex NP (*ComplexNP)*
What did Claire make the claim that she read in a book?
What did Paul hear the rumour that I found in my garage?
Where did you discover the fact that the criminals put the car?
Where did you put forward the hypothesis that all the weapons were?

*Negatives without do support (*Not)*
Her husband not claimed they asked where we were going.
The manager not implied you knew what they were doing.
The teacher not suspected she remembered where that woman lived.
Your sister not believed I forgot what he had done.

*Declaratives with double tense marking (*DoubleTn)*
His cousin doesn’t thinks we lied because we were afraid. The girl doesn’t remembers where she spent her summer holidays. The mother doesn’t knows Julia was absent from school today. Your brother doesn’t believes the man is telling the truth.
References

Ambridge, Ben and Adele E. Goldberg
2008 The effect of verb semantic class and verb frequency (entrenchment) on children’s and adults’ graded judgments of argument structure overgeneralization errors. *Cognition*.

Bradac, James J., Larry W. Martin, Norman D. Elliott and Charles H. Tardy

Culicover, Peter W.

Da ˛ browska, Ewa and Elena Lieven

Frazier, Lyn and Charles Clifton, Jr.

Goldberg, Adele E.
Haegeman, Liliane  

Hawkins, John A.  

Hiramatsu, Kazuko  

Holmes, V. M., L. Stowe and L. Cupples  

Kluender, Robert and Marta Kutas  

Lakoff, George  

Langacker, Ronald W.  


Levine, Robert D. and Thomas E. Hukari  

Lieven, Elena V., Heike Behrens, Jennifer Speares, and Michael Tomasello  

Moon, Rosamund  

Ouhalla, Jamal  

Sinclair, John (ed.)  

Snyder, William  

Spencer, N. J.  

Theakston, Anna L.  
Thompson, Sandra  

Tomasello, Michael, Nameera Akhtar, Kelly Dodson, and Laura Rekau  

Trueswell, John C., Michael K. Tanenhaus, and Christopher Kello  

Verhagen, Arie  

Wasow, Thomas, and Jennifer Arnold  