Questions after stories: Supplying context and removing it as a variable.

Jill G. de Villiers

Thomas Roeper

Follow this and additional works at: https://scholarworks.umass.edu/aae_groundwork
Dear Barbara,

Thank you for your request. I am happy to grant nonexclusive permission to upload “Questions after stories: Supplying context and removing it as a variable” to the UMass Amherst institutional repository ScholarWorks, for academic noncommercial purposes. Please credit the reprinted chapter to J. G de Villiers & T. Roeper, Questions after stories: Supplying context and removing it as a variable, in D. McDaniel, H. Cairns, & C. McKee (Eds.), Methods for Assessing Children's Syntax, ©1996 MIT, Cambridge MA: MIT Press, (pp. 163-188).

Please let me know if you have any questions,

Pamela

Pamela Quick
Subsidiary Rights Manager
The MIT Press
One Rogers Street
Cambridge, MA 02142, USA
1-617-253-0080

***

Barbara Pearson

From: Pamela L Quick <quik@mit.edu>
Sent: Saturday, March 9, 2019 6:33 PM
To: Barbara Pearson
Subject: Re: Permission to post to our institutional repository

Dear Barbara,

Thank you for your request. I am happy to grant nonexclusive permission to upload “Questions after stories: Supplying context and removing it as a variable” to the UMass Amherst institutional repository ScholarWorks, for academic noncommercial purposes. Please credit the reprinted chapter to J. G de Villiers & T. Roeper, Questions after stories: Supplying context and removing it as a variable, in D. McDaniel, H. Cairns, & C. McKee (Eds.), Methods for Assessing Children's Syntax, ©1996 MIT, Cambridge MA: MIT Press, (pp. 163-188).

Please let me know if you have any questions,

Pamela

Pamela Quick
Subsidiary Rights Manager
The MIT Press
One Rogers Street
Cambridge, MA 02142, USA
1-617-253-0080

***

Barbara Pearson

From: Barbara Pearson
Sent: Friday, March 08, 2019 8:59:49 PM (UTC-05:00) Eastern Time (US & Canada)
To: Hannah G Gotwals; Barbara Pearson
Subject: Permission to post to our institutional repository

Dear Ms. Gotwals,

I’m writing on behalf of Tom Roeper, a faculty member here at the University of Massachusetts Amherst. He would like permission to deposit his joint contribution (with Jill de Villiers of Smith College) to Methods for Assessing Children’s Syntax (full citation included below) to our institutional repository, ScholarWorks@UMassAmherst. As Project Manager and now co-author, I am organizing an archive (“collection”) of the project that led to the publication of the DELV, Diagnostic Evaluation of Language Variation tests. This chapter in McDaniel et al., 1996 is pivotal in the groundwork for the tests, and will go a long way to helping future scholars understand the conceptual framework for them.


Thank you in advance for any help you can be.

Barbara Pearson
(DELV Project Manager, and Co-author)

**********
Contents

Contributors ix
Preface xiii

PART I  1
Production Data

Chapter 1
Collecting Spontaneous Production Data 3

Chapter 2
Analyzing Children's Spontaneous Speech 23

Chapter 3
What Children Know about What They Say: Elicited Imitation as a Research Method for Assessing Children's Syntax 55

Chapter 4
Elicited Production 77

PART II
Comprehension Data 103

Chapter 5
The Intermodal Preferential Looking Paradigm: A Window onto Emerging Language Comprehension 105

Katherine Demuth
Karin Stromswold
Barbara Lust, Suzanne Flynn, and Claire Foley
Rosalind Thornton
Kathy Hirsh-Pasek and Roberta Michnick Golinkoff
picture showing the external reading (Goodluck and Terzi 1995). Thus, even judgment tasks are not always capable of verifying dispreferred readings (this issue is also discussed in chapters 10 and 11 of this volume).

Finally, in comparing the advantages of various tasks, it is important to keep in mind that for whatever reasons, one child may show his knowledge of grammar with task X and another child may show it with task Y. For example, in the study of Greek just mentioned, some children acted out external reference for the embedded subject of sentences such as (13) but rejected pictures showing such an interpretation; other children accepted such pictures but failed to show the external reference reading in their act-outs.

7.6 Conclusion

Act-out has a valuable place in the array of available tests of syntactic knowledge. Because act-out has been used extensively, its pitfalls are fairly well understood, as are its advantages. A body of act-out results now exists on a range of syntactic phenomena, providing a useful baseline for crosslinguistic work and for comparison with other techniques.

Notes

1. The reader who consults the paper just quoted will find that in Saddy's study of an adult aphasic patient, act-out produced a considerably lower level of performance (correct responses) than did a picture verification task. Saddy attributes this difference to the fact that the patient studied had a severe short-term memory deficit, which impaired his ability to plan the actions for an act-out more seriously than his ability to verify a picture as fitting a given sentence. Thus, the advantage of act-out over tasks such as picture verification in discouraging response bias and/or unreflective responses may be nullified by special characteristics of the population studied.

2. The subordinate clause was omitted with imperative stimuli approximately as often for before and after clauses in Johnson's study, even though, as Dana McDaniel (personal communication) points out, one might expect the subordinate clause to be omitted more frequently with before.

3. The design of act-out experiments was and is frequently sufficiently complex to limit the number of tokens of each sentence type presented to four or fewer, thus restricting the use of statistical remedies (in particular, treatment of items as a random effect) that Clark discusses as a solution to the problem of generalizing from the set of stimuli.

Chapter 8

Questions after Stories: Jill de Villiers and On Supplying Context and Thomas Roeper Eliminating It as a Variable

8.1 Introduction

In the philosophy of science considerable attention has been paid to the question of how to be sure that a given experimental result can be taken as supporting evidence for a theory. That is, theorists are prone to see the causal path in a rather straightforward manner:

Theory → Experimental result

If the experimental result occurs, they rejoice in the confirmation of the theory. But Duhem (1906), Quine (1953), and many others have written about the necessity of paying attention to the methods, measurement procedures, and even assumptions about the apparatus being made by the scientist in subjecting the theory to the test. That is, the result might have occurred not because the theory was true, but because other aspects of the experimental or methodological set-up made it likely to happen:

\[
\begin{align*}
\text{Methodology} \\
\text{Situation} \\
\text{Recording procedures} \\
\text{Theory}
\end{align*} \rightarrow \text{Experimental result}
\]

In other words, it must be recognized that every test of the theory is also a test of the "auxiliary assumptions."

Even more problematic is the notion of a confirmation of a theory by a result. Popper (1959) and his followers argued that only disconfirmation is scientific progress: the theory must predict that something will not occur, and if it does occur, then that theory can be proven wrong:

Theory → Not experimental result
For Popper, true progress in science came with the rejection of theories by such tests. Popper’s early work also discussed the importance of theories that made “risky predictions”—namely, ones that seemed counterintuitive or would be highly unlikely on the basis of common sense:

Theory \rightarrow \text{Not experimental result}
Common sense/intuition \rightarrow \text{Experimental result}

Much subsequent work raised doubts about achieving certainty by seeking disconfirming evidence, because it can be shown to involve the same problems of induction as positive evidence (Reichenbach 1978): namely, accumulation of compatible instances never guarantees the truth of the theory. Furthermore, negative evidence too is subject to the Duhem-Quine problem. The failure of the prediction might be due to some other assumption: maybe the apparatus was wrong, or the recording technique, or the environment for testing:

\[
\begin{align*}
\text{Methodology} \\
\text{Situation} \\
\text{Recording procedures} \\
\text{Theory}
\end{align*} \rightarrow \text{Not experimental result}
\]

In general, it is a combination of the evidence and the inherent logic of the theory that makes an experiment persuasive. In the absence of a set of theoretical commitments and beliefs, explanations for experimental results are considered ad hoc. In the most famous case of this problem, the experimenters (Michelson and Morley; see Michelson 1881) tried to test for the drag of ether on the speed of light by testing the speed of light at different angles to the earth’s motion through space. Ether was held to be the medium through which electromagnetic “waves” like light must travel. The assumption of the theory was that the apparatus was swimming in a stream of ether, and as with a swimmer attempting to swim in a river, the direction of flow should make a difference to the speed of light. When the expected result did not occur, and instead light seemed to travel at the same speed in both directions, Fitzgerald (1895) and Lorentz (1895) independently invoked an explanation involving the rather miraculous-sounding expansion and contraction of the arms of the apparatus, which was supposed to compensate exactly for the expected difference in speed, thus rescuing the theory about ether. The explanation remained ad hoc until Einstein’s theory of relativity provided the new paradigm to accommodate the results, in which the speed of light was constant.

Questions after Stories

How do we escape from this dilemma in our linguistic research? How can we know that a certain set of results is evidence on behalf of a theory and not due to the auxiliary assumptions required for the test? We have tried to do acquisition studies on aspects of reasonably refined linguistic theories, so that there are other kinds of evidence already articulated in their defense. In our experiments, the auxiliary assumptions are held constant across two conditions that elicit different responses, so the minimal difference between the two conditions must be held responsible for the responses. If the auxiliary assumptions are appealed to as explanation for one phenomenon, they can be shown to make precisely the wrong predictions for a second phenomenon within the same experiment.

We have described this as our ideal strategy; as researchers, we have undoubtedly slipped off this bandwagon of ideal methodology and at times placed greater emphasis on theoretical coherence than methodological rigor. Let us illustrate the strategy in its purest form, and then describe its history.

Suppose there is a sentence, S, that is subject to two interpretations; it is inherently ambiguous. Suppose further that it is preceded by some context that makes either interpretation readily available. One might then expect that subjects hearing this sentence would interpret it either one way or the other, depending on their personal predilections, how they saw the salience of the events in the context, and so on. Ideally, given 10 such sentence/context pairs, they might interpret S one way and S the other way. Even more ideally, if we looked at the behavior of 10 subjects, it would turn out that each story was interpreted one way five times and the other way five times. Given that result, it would be hard to claim either that individual stories contained biases or that individual subjects had parsing biases. We take this as proof that the stories are pragmatically balanced.

Suppose next that there is a sentence, S’, that is also subject to two interpretations. However, one of the potential interpretations is forbidden by linguistic theory; it would violate some universal constraint. The context preceding S’ is precisely the same as the context preceding S; it contains the same readily available interpretations. However, when the same subjects are presented with S’, they find only one interpretation of the sentence. The scenario has been constructed then in such a way as to rule out the contribution of context to the change in interpretation; since the context stays the same, it is implausible to argue that it is responsible for the shift in choice of interpretation.
Of course, in actuality, the same subject would not be given the same tokens of S and S’. Perhaps for one group of subjects S would be given for one set of five contexts and S’ for the other, and for a second group S’ would be given for the first set and S for the second. The groups would then be compared to ensure that the sets of tokens made no difference to the outcome.

This design underlies the first in a series of studies we have conducted in the domain of wh-question acquisition with young children. In the next section we trace the somewhat mottled history by which we achieved this design.

8.2 Putting It into Practice: History of the Procedure

Consider what is involved in the real-world comprehension of a complex sentence.

(1) How did John ask who to paint?

The situation semantics require at least two characters, some indecision on the part of one of them, and the need for two questions, one indirect.

The result is a situation that cannot be set up in a single sentence; rather, a story is necessary if the referential domain is to be included. Now, however, we have introduced all the problems of memory. Will the child retain the story as we wrote it? In fact, adults are sometimes boggled by a story if there is no visual context to which it can be anchored.

The presentation pragmatics of the experiment now become relevant. We need pictorial support for a child to remember the story that contains the context, but if we present the pictures to children in a sequential manner, perhaps they will pay attention only to the last picture. Should we then create a guarantee that they survey the entire array?

Finally, discourse pragmatics are relevant. Should we ask the children questions to which they know that we know the answers? Is it not pragmatically odd to ask for information that the questioner already possesses? Does the fact that parents constantly ask “test” questions mean that most children will be quite used to this situation?

These are not only possible questions—they are questions that we addressed very seriously in the first phase of our research when we were unsure that we could elicit responses at all to complex questions. In effect, we have introduced a totally uncontrollable set of variables were one to evaluate our work in terms of the variables themselves. Should we independently determine the nature of each of these variables before we begin research? We argue instead that if we design the study in the right way,

Questions after Stories

these methodological issues should assume a lesser importance. How? We use the classic logic of minimal pairs. All other things being equal, any response difference must be due to the theoretically motivated difference between the two sentences.

In our first study on wh-questions (de Villiers, Roeper, and Vainikka 1990) we presented children with stories followed by ambiguous questions that permitted the children a choice between two interpretations. For example, children might hear the following short story (accompanied by pictures like those in figure 8.1):

Figure 8.1
Typical pictured story from de Villiers, Roeper, and Vainikka 1990
Story
This boy loved to climb trees in the forest. One afternoon he slipped and fell to the ground. He picked himself up and went home. That night when he had a bath, he found a big bruise on his arm. He said to his dad, "I must have hurt myself when I fell this afternoon!"

Question
(2) When did the boy say he hurt himself?

Given the story, the question has two possible interpretations, depending on where the listener interprets the trace to be for the wh-question when. Is it connected as an adjunct to say, as in (3), or to hurt, as in (4)?

(3) When did he say t1 he hurt himself?

(4) When did he say he hurt himself t1?

That is, a subject might answer "At night" (if (3)) or "That afternoon" (if (4)), depending on the interpretation. The answer corresponding to (3) is referred to as a short-distance movement, because the wh-word moves within the first clause (from the position marked as the trace t1). By contrast, the answer to (4) involves long-distance movement, because the wh-word moves from a position in the embedded clause (t1).

For that study, we designed stories containing information that made both possible interpretations salient, but we were worried that given the choice, children would choose the short-distance interpretation. Hence, we introduced a source of bias in the context: we made a point of having a second "stooge" experimenter who secretly showed the child a picture containing the information needed to answer the long-distance question from the first experimenter, hoping in that way to increase its salience and plausibility as an answer. For example, for the question in (5)

(5) How did Big Bird ask to help?

the "stooge" experimenter showed the child (but not the questioning experimenter) a picture in which Big Bird was helping by cutting out the cookie dough—in other words, a picture that provides the "long-distance" answer about how Big Bird helped. The 3- to 6-year-olds (de Villiers, Roepers, and Vainikka 1990) were quite happy to provide either answer, suggesting that their grammars do readily permit long-distance movement. We dropped these extra features in later work when we discovered that they made no difference in the results obtained. And, as is often the case, the evolution of the experiment itself pointed to progressive simplifications.

The main point of the study, however, was to test whether children in this age range obey the subtle constraints on movement that are so central to contemporary syntactic theory. In long-distance questions, the wh-word moves in two steps, so if another wh-word is present, the first move is blocked; in contemporary terms, there is a barrier to movement (Chomsky 1986a). If the medial position in a question is already filled by a wh-word serving as the complementizer (the embedding connective), a wh-question cannot cycle through that position and leave a trace (t1), so the resulting sentence is ungrammatical.

(6) *When, did he say how he hurt himself t1?

The long-distance interpretation 'in the afternoon' was all right for (4), but it is not permissible as an answer for (6). The question in (6) sounds grammatical, but only under the interpretation of 'when did he say it', not 'when did he hurt himself'; the long-distance reading is excluded. The concept of local movement is a feature of Universal Grammar; hence, a child should have that option available. We controlled the stimuli so that for any given story, half of the children heard a story followed by a sentence without a medial wh-word, and the other half heard the same story followed by a sentence with a medial wh-word. By designing the study this way, we approached some of the ideals discussed above. First, we had a theory that forbids certain things to happen. If the child did permit long-distance movement in the condition forbidden by the theory, the child's response would be strong evidence against that theory of grammar for young children. However, if a theory survives such a test, it is not proven by it; it merely remains in the set of possible theories. Second, the theory suggests a result that is counterintuitive: adults naive to linguistics are always surprised by the contrast between (4) and (6) and usually resort to ad hoc semantic explanations once they notice it. Third, because the method, context, and measurement are identical, any difference between performance on (4) and performance on (6) must be a consequence of the subtle difference between the two sentences, namely, a medial wh-word occupying the specifier of CP position.

The findings of that study revealed that 3-year-old children know the Universal Grammar principles, in that they only rarely give long-distance answers to questions such as (6), which contains the medial wh-word. There was a consistent, often tenfold difference between the percentages
of long-distance interpretations for questions with and without the barrier. There were further subtle outcomes such as a contrast between adjunct (e.g., how, why, when) and argument (e.g., what, who) questions that is also strongly counter to lay expectations but precisely predicted by contemporary formulations of barrier theory (e.g., Chomsky 1986a; Rizzi 1990; Cinque 1990). We were not the first to use the story-and-ambiguous-question technique; earlier projects exploited the same procedure. Chomsky (1969) used ambiguous questions; Maratsos (1976) used stories to explore sensitivity to subtle contrasts in determiners. Others put the two techniques together. Otsu (1981) was among the first to use the combined technique, to test directly whether children obey constraints on wh-extraction from relative clauses and prepositional phrases in a comprehension task. His preschool subjects heard short pictured stories followed by a wh-question. For example:

**Story**

Jane is drawing a monkey with a crayon. The monkey is drinking milk with a straw.

**Question**

(7) What is Jane drawing a monkey that is drinking milk with?

Adult speakers of English can find only one possible meaning for the question, namely, one that links the question to the main clause and construes the preposition as being associated with that clause.

(8) What, is Jane drawing [a monkey that is drinking milk] with that?

However, notice that the question is potentially ambiguous, if the child lacks the constraint on extraction from a relative clause: it could mean "What is the monkey using to drink the milk?" if the trace of what is construed as inside the relative clause.

(9) What, is Jane drawing [a monkey that is drinking milk with that]?

If the child’s grammar had no constraints on extraction, the answer “A straw” would reveal the availability of the embedded site. In addition to relative clause sentences, Otsu tested the children on simple questions with prepositional phrases. These questions followed simple pictured stories.

**Story**

James is painting a picture of a boy with a book. James is painting a picture of the boy with a brush.

**Question**

(10) What is James painting a picture of a boy with?

Actually, Otsu’s results showed that 3- and 4-year-olds did rather poorly on this task, often answering the questions in a manner that violated the barrierhood of the relative clause or the prepositional phrase. However, he also included a test of comprehension of the relative clause, which revealed the very interesting finding that those children who showed mastery of the structure in comprehension where more likely to respect the barrierhood of the relative clause. Some of the reasons for the relatively poor performance of children on Otsu’s experiment are discussed in de Villiers and Roeper 1995; we found striking obedience when adjunct questions were used instead. Notice that Otsu’s procedure, although consisting of the same general idea of a story followed by an ambiguous question, has no built-in control for the influence of the story on the interpretation.

Phinney (1981) also used the ambiguous-question technique to explore children’s understanding of that-trace restrictions in English. Her task involved short pictured stories with two potential referents for a question such as (11) that would be ambiguous if the child lacked the That-Trace Filter.

(11) Who did the lion know that swam in the pond?

If this question is construed as having a trace following that—

(12) Who did the lion know that [that swam in the pond]?

in other words, who did the lion witness swimming in the pond—it violates the That-Trace Filter. However, the child had the alternative of choosing a relative clause reading for the sentence.

(13) Who did the lion know [that swam in the pond]?

In this case, the lion knew someone, who swam in the pond unbeknownst to the lion. The pictured story defined different referents for the two readings, but the children only rarely chose the reading that would correspond to a violation; that is, they didn’t choose the creature witnessed swimming. Phinney’s evidence could therefore be considered in keeping with the notion that English-speaking children obey the That-Trace Filter. Notice that, like Otsu’s, this procedure lacks the control against bias in the story favoring one interpretation. But in Phinney’s case subjects did choose the witnessed swimmer in answer to the question lacking that.
Questions after Stories

(14) Who did the lion know swam in the pond?
The work of Otsu and Phinney thus tested precise, theoretically important distinctions by the use of potentially ambiguous questions.

8.3 Assumptions and Controls

In this section we discuss some of the underlying requirements of the task so that experimenters who wish to use this method to test their own theoretical questions can decide whether it is appropriate for that purpose. What are we assuming about the cognitive and linguistic processes in the task?

The child has several problems to solve. First, she must understand the story: the characters, their motives, the order of events, the paths of causation. Information for this is coming from two sources: the story being read, and the sequence of pictures laid down in order as the story is read. The pictures are snapshots capturing critical events; time and motions are lost, and presumably important clues given in real life by facial expression and nonverbal cues from participants are captured only grossly by the artist. Nevertheless, we assume that the average 3-year-old has some practice in this art of narrative construction from being read to both at home and at school.

Second, the child must retain enough of the narrative in memory to search for the answer to the question asked. Again, it is presumptuous to conclude that 3-year-olds have exactly the same skill level in this regard as 5- or 6-year-olds; however, we try to keep the stories short enough to be within their grasp, and of course the pictures are still present as recall clues.

Third, the child must parse the question and arrive at an interpretation for the site of the trace. In our theoretical work, we tend to talk as though that is the only task involved. But how does it interact with the other two tasks above?

Consider a question such as (15).

(15) How did the mother say she cooked the pie?

One possibility is that sentence processing is extremely modular. On that scenario, the child hears the question as an entirely isolated piece of language, and identifies the trace location and arrives at an interpretation in complete independence from the story or pictures that have just been presented. Suppose that this process identifies the first possible gap (after say) as the site of the trace. Having arrived at an interpretation, the child must then look at her memory representation to find the answer to, for example, How did the mother speak?, and having found such an answer, she gives it.

Consider a second possibility, which is that the processing of the question is completely permeable to the context created by the story. On that description, the child might have already paid greater attention in the story to the manner of speaking, say, than to the manner of cooking. When she hears the question, she immediately construes it to be about the act of speaking rather than about cooking, and again searches her memory representation (presumably it is already primed) and retrieves the answer.

A third position is a compromise between these two: perhaps the sentence interpretation is encapsulated, but in the search process one manner is much more salient than the other and the child might backtrack to a different interpretation if the first interpretation offers no immediate answer.

These are important alternatives to be pursued, but we have not yet spent research time in exploring them. Is this a shocking neglect? No, for the following reason. Whatever the process is, it will have to be common both to sentences with a barrier to wh-movement and to sentences without. The presence of, say, a medial wh-word, or an adverb, in the final question cannot plausibly change the salience of events already coded from the story, or the accessibility of the events from memory. There must be a parsing difference, but that is a reflection of the grammatical constraint, and any theory must accommodate that. Hence, any systematic differences that we detect between these sentences cannot be an artifact of the kind of story processing being undertaken, but must instead reflect the child's grammatical parsing alone.

Why, then, do we attend to the design of the stories so carefully? We do so because it is our goal to minimize the memory and comprehension demands for young children so that failures to arrive at an answer do not confound the results. Remember that no child sees a story twice. The child who hears story A followed by a question with no barrier will hear story B followed by a question with a barrier. Another child will hear story A followed by the question with a barrier and story B followed by the question without one. If we have ensured the comparability of pairs of stories with and without a barrier, then differences that we obtain can be attributed to knowledge of the barrier. In other words, we have achieved
the goal we set out with: we have supplied context and yet removed it as a variable.

There is one final source of difference that it is natural to attach importance to: the prosody of the question. Much has been written about the important role of sentence prosody as a bootstrap to sentence parsing for young children, though the case is presently built from rather indirect evidence. If it is true that children are highly sensitive to the intonation pattern across a sentence for providing major clues to, say, the location of gaps or the coindexation of anaphora, then it is essential to control that variable in presenting matched pairs of questions.

The usual judgment for many adults is that a speaker can bias a two-clause question to a lower-clause interpretation by using a “long rise” type of intonation with no special stress on the matrix verb.

(16) When did she say she ate?

Alternatively, a speaker can bias the question to a short-distance interpretation by using a slightly falling intonation and a light emphasis on the matrix verb.

(17) When did she say she ate?

In our own work we have tried to control intonation as reliably as possible. Nevertheless, experimenters are human, and the vagaries and stress of presenting 50 or 60 stories and questions over the course of a morning mean inevitable variation in prosody. Fortunately, we videotape our sessions and have all the questions stored on tape. Last year a research assistant blind to the hypothesis at hand went back and coded a random sample of 180 questions for their prosodic characteristics: either long rise or fall or unclear/neutral. The children’s responses to those questions were then analyzed to see if the prosody influenced their choice of answer. Results were as follows: the sample included 65 ambiguous wh-questions with no medial barrier, of which 36 were presented with rising intonation, 24 with falling intonation, and 5 with neutral intonation. Figure 8.2 shows the distribution of answers to those questions, and it does seem that long-distance answers are more readily provided under the long-rise prosody. The sample also included 168 questions that involved wh-barriers, and here the distribution of prosody was different: 35 questions were produced with rising intonation, 87 with falling intonation, and 48 with neutral intonation. Apparently the experimenter found it hard to produce the long-rise prosody reliably with a medial barrier. Figure 8.3 shows the distribution of responses to those questions: the children resist long-distance answers appropriately, and do so even for the long-rise intonation.

Given the less than systematic presentation of the prosody, the intuition about prosody needed further exploration. Another experiment was therefore conducted in which prosody was systematically varied. Children were presented with eight stories followed by questions, half of which had long-rise prosody and half of which did not. The stories were presented in the usual manner. There were two sets: in one set a group of children heard half the questions with a medial wh-barrier, in the other they heard those questions without a barrier.

The results were discouraging for the idea that prosody exerts a strong influence. In fact, they were unexpected: the falling intonation actually marginally encouraged long-distance answers in the case of questions with no barriers. In the case of questions with barriers, prosody had no effect. Given that the two studies were conducted by different experimenters, perhaps prosody is a feature that varies idiosyncratically. More work comparing adults and children might reveal interesting changes. Nevertheless, the results from both studies suggest that children can resist long-distance answers that would violate barriers even in the face of misleading
Questions after Stories

not to arguments such as who or what. For (21), it is easy to get a long-distance reading.

(21) What did he always say he won?

The details of the theories devised to account for this matter do not concern us here. The main point is, How might one investigate whether children know weak barriers?²

Philip and de Villiers (1992) set out to investigate just this question with respect to adverbs of quantification such as always. They sought to find out whether children would block an interpretation of an adjunct question from a lower clause if an adverb of quantification intervened. To limit other sources of variability, they focused on a single adjunct question, why. They did this because the other candidates were less suitable: when would interfere with the use of an adverb such as always in awkward ways.

(22) When did he always say he won?

However, since there was not much previous work on why as a long-distance question, they needed to be sure that children would readily give long-distance and short-distance answers to why questions if no barrier was present. This need led to part of the design: roughly identical stories would be constructed that contained answers both to the question "Why did someone say something happened?" and to the question "why did something happen?" For one of them, the question would be simply as shown in (23), and for the other, a quantificational adverb would be added, as in (24).

(23) Why did x say y happened?

(24) Why did x always say y happened?

The predictions would then be as follows: For (23), if the stories were constructed to be free of bias, the children should be equally prone to answer with long- or short-distance interpretations. For (24), even though the children heard the same stories, the question should lead to only short-distance answers if children understood the barrierhood of the adverb. By comparing answers to (23) and (24) for the same children, it would be possible to assess the degree of obedience to the barrierhood of adverbs.

The goal, then, was to devise stories that had the right properties: equal likelihood that a short- or long-distance answer would be given if the
question were of type (23). The events in the stories had to be balanced so that the reasons for speaking and the reasons for the event spoken about were equally salient. The rationale for undertaking this kind of balancing has already been discussed. However, here the design was counterbalanced to test the influence of position and type of connective on biasing children's answers to the ambiguous questions (i.e., type (23)).

The design should not allow one type of reason to be lost in the middle of the story and the other to be either highlighted by being at the end or subject to undue emphasis. For example, if the reason for speaking was always embedded in the middle of the story, but the reason for the event was always the last thing mentioned, the child's memory might be biased in such a way that the last reason mentioned was always the first reason to pop into mind. Having no evidence either way on such a matter, Philip and de Villiers chose to counterbalance where the reason occurred within the story and test position as a factor influencing the answers given to type (23) questions.

It was also of concern that some types of causal connection might be more familiar than others to young children. Previous work on naturalistic production (e.g., Bloom et al. 1980) had discovered that different connectives appear at different times in children's speech, so the reasons should not be confounded with types of connective. The connectives chosen were because and so that, and the stories were balanced in such a way that the type of reason was introduced by either connective. Once again, stories were counterbalanced to test the influence of type of connective on the answers to type (23) questions.

In fact, no one knows whether such biases really exist, and this issue has not been explored to date except in the study just described. In all other cases, we have taken care to balance the stories or to disentangle the variable of interest from other extraneous influences such as order of mention or salience. It may be impossible to foresee every potential confound. Why should certain factors, such as intonation or order of mention, always be considered more likely candidates for explanation, though highly ad hoc, than the variables derived from serious theoretical work? Philip and de Villiers (1992) decided it may be more productive to confront the variables head-on and test for effects than to try to balance all kinds of potential confounds. This approach makes the results more plausible, but since the set of potential pragmatic variables is infinite, ultimately reliance on the intrinsic strength and coherence of the theory is unavoidable.

Questions after Stories

In the end, Philip and de Villiers controlled intonation, controlled salience, and counterbalanced position in the story and connective. The latter two variables proved to be insignificant variables in the children's choice of answer for sentences of type (23) when tested in an analysis of variance.

Each child heard 4 stories followed by questions of type (23) and 4 followed by questions of type (24) containing an adverb. Four cases were also included with strong medial-wh barriers like (18). Each child thus received 12 example stories and questions, usually in two sessions of about 10 minutes each. (In this study the children were asked to complete another task as well.) As in all our work, the stories, were randomized anew for each child (by shuffling the set) so they were not presented in the same order. The remaining parts of this experimental set-up were sufficiently general that we highlight them in a separate section.

8.4.2 More General Procedures

A further aspect of the design concerns the pictures that accompany the stories. We endeavor to have a layout that consists of at most five pictures, because it is hard to scan across a larger set. Usually a simple story can be captured with three pictures that depict the most salient events. Again, the pictures should ideally depict both events discussed, without being biased to favor one reading rather than another. We have used a variety of artists and find that cartoonists, provided their style is not outlandish, produce pictures more suited to the task than other kinds of skilled artists who may pay too much attention to the aesthetics and not enough to what is salient in the story. We have used both black-and-white and colored pictures. In most cases the backgrounds are very minimal and the activities are the focus. An example was provided in figure 8.1.

For most of our studies, we pilot: test our stories and pictures first on college students, who are tested in (almost) the same way as the children. On occasion we have given these as group tests using an overhead projector, but it has been our experience that college students sometimes take these tests as opportunities for having fun if they write the answers and are assured anonymity! We get more sober and reasonable answers with one-on-one interaction in which we can explain the purpose of collecting adults' data for comparison with young children's data.

Finally, there is the manner in which we interact with the child in the sessions. It is imperative to have a quiet place to record the session free of interruption. We first welcome the child informally and introduce...
ourselves, then draw attention to the fact that we are recording on a video camera by asking the child to “watch for the blinking light to make sure it’s working.” It is preferable to make this explicit and then ignore the camera rather than adjust the focus (or whatever) without talking to the child about it. The child is seated at a small table on which we can lay out the pictures relevant to the story. It is important to videotape the sessions to record nonverbal responses (e.g., pointing to a picture).

We ask the child if she likes to listen to stories, and say that we will read a story and then ask a question about it. As the story is read, one experimenter lays out the relevant pictures from left to right in front of the child, leaving them all visible at the time of questioning. At the end of the story, the experimenter pauses briefly, then asks the question. A second experimenter records the child’s answer verbatim and later checks it against the videotape.

If a child hesitates, acts shy, or looks bewildered, the first experimenter repeats the question. If met with the same response, the experimenter asks if the child would like to hear the story one more time, then repeats it. If a child just points to a picture, the experimenter attempts further clarification by asking, for example, “What? Can you tell me instead?” Once a question is answered, the experimenter proceeds to the next. Children of this age are prone to interrupt with comments or questions of their own, and we do not discourage natural conversation of this sort. If the interruption has been prolonged or off topic, we typically start the story from the beginning after it. With stories this brief, we find most children reserve their conversation for the pause between stories.

Children’s pleasure in doing the tasks is relevant; we take it to mean that they are engaged and using their abilities. When the task has been completed, we thank the child profusely and regardless of performance we make some congratulatory comment such as “How come you know so much stuff already? I bet you’re really six years old!” The purpose of this is to send the child back to the classroom with a feeling of accomplishment and goodwill. We do not lack for volunteers the next day (of course, we are restricted to choose from the children who have permission to participate).

Data analysis is the next task. We transcribe all responses and enter them in a relational database (4th Dimension) on the computer, with separate files for information on the children (age, gender, school), the stories (type, question), and the data (actual statement by the child, then codes).
Using the database allows us to sort along many variables, for instance, to print a report of all the answers to one particular story, or all the “short-distance” answers from 4-year-olds. It is not a necessary step but it is one that has illuminated several phenomena that might have stayed hidden in the original score sheets. In addition, the data set can still be coded along new dimensions of interest. As just one example, we became interested in whether children were prone to answer with an appropriate constituent or not. Consider question (26).

(26) How did the mother say the boy rode the bike?
Let’s suppose that in the story the mother said it on the telephone, and the boy rode the bike through the mud. For the upper clause, a child could give the answers in (27a), and for the lower clause, the answers in (27b).

   “By telephone.”
   “That.” (pointing)
   “She spoke in the telephone.”

   b. “Through mud.”
   “Mud.”
   “He rode it through mud.”

We coded the constituent form of the answer (e.g., PP, N, NP, S) to test the hypothesis that the occasional mis-answer (e.g., a violation of a barrier) was less likely to be a constituent than a correct answer was, because that might suggest the child was retrieving the answer by some means that circumvented grammar and used inference. We have no strong evidence yet on behalf of that idea, but the database allowed us to recode data along a new dimension to answer a new theoretical question.

After the data are entered, we make a new data sheet suitable for analysis by a statistical package (e.g., analysis of variance). Philip and de Villiers 1992 were interested in the number of lower-clause answers to questions of type (23) without the adverb, versus type (24) with the adverb. Therefore, for each child the number of lower-clause answers to each type of question was counted. Since the contribution of position in the story and connective was found to be insignificant, those differences were collapsed in the final analysis. In other studies such variables as gender and age or school might be of interest, and these would be entered, as between-subjects variables in the analysis of variance.

Questions after Stories

In the Philip and de Villiers study, the weak barriers created by adverbs were not respected by 4- and 5-year-old children. The children were just as likely to give a lower-clause reading as an upper-clause reading regardless of whether a quantificational adverb was present in the question. The same children were tested on “strong” barriers—medial-wh-islands—and respected them completely, suggesting that different mechanisms or modules are at work in the two cases (see Philip and de Villiers 1992).

8.4.3 When Is the Method Most Useful?
The technique as described has been used most successfully in, and is perhaps best suited to, the following kinds of circumstances:

- If one can test children on matched pairs of sentences (e.g., one with a barrier and one without) and control all the other features. However, it may also be useful for a contrasting pair of sentences for which, if a child knows the contrast, one answer would be given for one question and the other for the other question. At present it seems best suited to studying wh-interrogatives, but it might be extended with some imagination to other constructions.
- If children can be given enough examples to allow robust statistical analysis: at least four examples of each type, six or more if individual grammars are of interest rather than group data.
- If the materials are suited to age 3½ and upward. Below that age we have encountered difficulties with getting children to give full answers rather than pointing; we have also found that children have trouble concentrating on the story and tend to lack the basic world knowledge and lexical and syntactic proficiency to follow the events and the question.
- If the events are picturable and do not involve subtle issues of intentionality and theory of mind, which confound the picture of syntactic development in this age range (for examples, see de Villiers 1995).

These limitations have not stopped us from pursuing other questions, however. For example, we have attempted a longitudinal study to explore how individual grammars of wh-movement change over time. Working in a cooperative campus nursery school, we tested the same children over the course of one year on a variety of wh-question forms, to see if we could follow how their grammars changed (see de Villiers and Roeder 1995, for a partial report). We learned from this experience how important it is to present enough exemplars of a type in any one testing session, but were
driven by a rival desire to get as much information on as wide a variety of phenomena as possible in any one round of testing. As a result, the goal of diagnosing a particular child's grammar at any one point is compromised by the small number (usually two examples) of a type of sentence presented at any one time. In addition, since we had to make up new examples every three months, the stories and pictures were unevenly matched and some of the fluctuations in performance may be due to the materials rather than to real growth or regression. Obviously, the lessons learned were valuable ones: we realized the importance of selecting, balancing, and pilot-testing the materials in advance for each round. In addition, we recognized that to use this procedure to "diagnose" a child's grammar of questions would require presenting probably six examples of each given type, to allow us to set criteria on mastery at one moment in time. If a child answered two out of only two questions with short-distance answers, could we conclude that the did not permit long-distance answers? If he answered one of the two "barrier" questions with a long-distance answer, could we conclude that he did not know the barrier effect? The more examples, the more secure these decisions can be for an individual child. Yet the limit for testing most children around age 3 seems to be about 15–20 short stories in a session, which can take only 10 minutes or so. Careful planning is necessary if enough data are to be gathered to be useful at the individual level.

Despite the shortcomings, we discovered some phenomena of interest through this longitudinal study and would recommend it again to others who had the opportunity to test changes over time. One interesting methodological note: children of this age remember pictures with astonishing accuracy. In one story we reused part of a set of pictures that a child had seen one year previously, and the child said, "I've already seen this one." We do not know if children would remember their answers, but it is a possibility to be beware of.

8.4.4 Pitfalls
Before we consider the pitfalls of the technique, let us mention some rather specific difficulties that have concerned us in the work on wh-questions. Confusions of meaning among the wh-words make it difficult to interpret some answers from the youngest group. Consider question (28).

(28) When did the boy say he hurt himself?

Questions after Stories

If a child answers this question by saying, "On the tree," can we be certain how to analyze the same child's answer to (29)?

(29) When did the boy say how he hurt himself?

If the child again says, "On the tree," is this phrase a long-distance answer to when, with a confusion of meaning, or an answer to the medial how?

The most prevalent answer that created coding difficulties involved a confusion of meaning for the adjunct question how, which may relate to an interesting theoretical perspective on adjuncthood (see de Villiers 1991). Children were very prone to interpret how as why, either by giving an answer that referred to a reason rather than a manner, or by answering, "Because." Hearing the question in (30),

(30) How did the dog run?

children often answered, "Because he stole the bone" (the reason) or (less often) "Because he was fast" (the manner). We do not know why this particular direction of confusion occurred; perhaps how questions are asked of children less often than why questions, or perhaps the semantics of how are harder to grasp. The control for this is to ask some simple questions at the beginning of the session that might reveal whether the child has mastered the meanings of the wh-words. We did this in the longitudinal study mentioned above, at concomitant sacrifice of attention to enough "interesting" examples.

The verb in the question has well-known effects on the youngest children's question answering. Children have been shown to be influenced by the probability that certain verbs will occur in certain wh-questions (Tyack and Ingram 1977; Winzemer 1981), and they will answer the "expected" question. For example, where is a likely question in conjunction with the verb go, so a young child who is asked, "When did they go?" is quite likely to mis-answer, "To the movies." Of course, controlling for this problem introduces new design questions.

A similar difficulty plagued us with what we came to call quotation answers. Let us suppose the question is as follows:

(31) How did the mom say she stirred the cake?

If the child pipes up in a squeaky voice, "I stirred it with a spoon!", how should that be scored? The long-distance answer is contained within the quoted reply, but by mimicking the mother's voice, is the child not telling how the mother said it? We ended up creating a special category for such
answers; fortunately, they were not especially frequent and did not seem to occur only in one particular category (e.g., only in barrier-type sentences). The issue remains for future work but points out again the difficulty with particular adjunct questions such as how.

When questions create difficulties of their own, because children’s grasp of time at this age is rather fragile, as may be their memory for order of events. However, their marking of major periods of the day (night, afternoon, next morning, etc.) may be more dependable, as may their understanding of the names of various times of year (spring, summer, winter, fall, on his birthday) or the moment of events (at the start of the race, when he climbed the tree, after she got home). However, 3-year-olds do not seem very accustomed to being asked when questions, and they sometimes resort to pointing at the pictured events, which is ambiguous in itself. The experimenter can ask for clarification (e.g., “Do you mean then?”), but it is hard to do so without leading the witness. Our solution has been to doubly mark events in the story for time (e.g., by saying, “when she got home that night” or “in the morning before she got up”), thus providing children with a richer context and more opportunities to refer to the time in whatever manner they prefer.

However, these particular difficulties concern not so much the story-and-ambiguous-question technique per se as the domain of wh-questions and their interpretation. A more important criticism can be leveled at the method itself: can we be sure that a child is “avoiding” the answer that would violate adult grammar? We have tried to argue that we can have confidence in this conclusion in the ideal situation described: if the child does not avoid such an answer in the absence of a barrier or constraint, and given a precisely equivalent story. The result should probably be backed up with other methods: judgment tasks and elicited production have been used for this purpose by others (McDaniel, Chiu, and Maxfield 1995; Thornton 1990) and have produced similar results.

8.5 Conclusion

We have argued that if one builds an experiment around a contrast that constitutes a minimal pair, then one can risk the use of an intricate context, that is, a story and accompanying pictures. If the results show a clear contrast, as they do, then the potentially “uncontrolled” features of the context have no consequence.

Questions after Stories

This does not argue that traditional methodological concerns are irrelevant. In fact, as described here, we use these practical insights about experimental design in attempting to balance our stories. Should another researcher decide that the results are due to a methodological feature not accommodated by the linguistic theory, then it is that researcher’s obligation to elevate the methodological feature into a countertheory. Such a countertheory must then meet the requirements of a theory: it must predict all the results. On this view, all experiments contrast rival theories; however, rival theories can be compared only if they make predictions of equivalent precision. It is, of course, not possible for us to articulate all rival theories, and new linguistic theories will surely be developed that may cast a different light on these results.

Notes

1. Together with colleagues we have explored children’s knowledge of these barriers in other wh-movement languages: German and French (Weissenborn, Rooper, and de Villiers 1991), African American English (Seymour et al. 1992), Caribbean Spanish (Pérez-Leroux 1991), and Greek (Letieri 1991). Kudra, Goodluck, and Progovac (1994) have also reported obedience to barriers in child speakers of Serbo-Croatian. Young children hearing translations of the test sentences into their own languages have provided convincing evidence of long-distance movement, barrier effects, and the argument/adjunct distinction (Rooper and de Villiers 1994; Maxfield and Plunkett 1991; see also Goodluck, Sedivy, and Foley 1989, and for application to the study of parasitic gaps, Maxfield and McDaniel 1991).

2. Theories of the phenomenon vary and include both syntactic and semantic accounts (Rizzi 1990; Szabolcsi and Zwarts 1990).

3. This unexpected result has suggested the interesting hypothesis that children fix the wh-feature prior to learning the full meaning of the lexical item.
References


References


Chien, Y.-C., and K. Waxler. 1990. Children’s knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. Language Acquisition 1, 225–295.


References


Chomsky, N. 1994. Bare phrase structure MIT Occasional Papers in Linguistics 5. MITWPL, Department of Linguistics and Philosophy, MIT.


References


References


References


Gordon, P. 1982. The acquisition of syntactic categories: The case of the count/mass distinction. Doctoral dissertation, Department of Psychology, MIT.


Jakubowicz, C. In preparation. Comprehension and production of complement clitics by Spanish-speaking children. Ms., Laboratoire de Psychologie Expérimentale, CNRS, Université de Paris V.


References


References


References


References


McNeill, D., and N. McNeill. 1966. What does a child mean when he says “no”?


Neville, H., D. Mills, and D. Lawson. 1992 Fractionating language: Different neural subsystems with different sensitive periods. Cerebral Cortex 2, 244–258.


References


References


References


References


