Memory & Cognition
1993, 21 (4), 477-487

A case study of anomaly detection: Shallow semantic processing and cohesion establishment

STEPHEN B. BARTON and ANTHONY J. SANFORD
University of Glasgow, Glasgow, Scotland
and ESRC Human Communication Research Center
Universities of Edinburgh and Glasgow, Scotland

Although the establishment of a coherent mental representation depends on semantic analysis, such analysis is not necessarily complete. This is illustrated by failures to notice the anomaly in questions such as, "When an airplane crashes, where should the survivors be buried?" Four experiments were carried out to extend knowledge of what determines the incidental detection of the critical item. Detection is a function of the goodness of global fit of the item (Experiments 1 and 2) and the extent to which the scenario predicts the item (Experiment 3). Global good fit appears to result in shallow processing of details. In Experiment 4, it is shown that if satisfactory coherence can be established without detailed semantic analysis, through the recruitment of suitable information from a sentence, then processing is indeed shallow. The studies also show that a text is not understood by first producing a local semantic representation and then incorporating this into a global model, and that semantic processing is not strictly incremental.

The idea that text comprehension is based on the construction of a coherent mental representation of what is in the text may be said to constitute an orthodox view. Essentially, coherence requires that there are no logical or semantic contradictions in the representation. This leads to a position in which coherence establishment depends on a thorough check of the assignment of fillers to roles. So, in comprehending the sentence *Ellen ate some steak for dinner*, the term *steak* is assumed to fill the object slot of something that was eaten for dinner. The acceptability of steak as a role filler depends on some process of matching its meaning to selection restrictions on the role slot. By such tests, the sentence *Mary ate some rocks for dinner* is discovered to be anomalous, because the term *rocks* fails to meet the criterion *is edible*, and as a consequence, local incoherence in the text is discovered, which would lead the reader to want to know just what was meant by the statement. The apparent ease with which such anomalies are detected suggests that a thorough check is made of filler against role specification as part of the normal process of coherence establishment. Recently, the completeness of such checking processes has been called into question. In the present paper, we shall investigate this issue.

In addition to completeness of testing for the fit of fillers to roles, a commonplace assumption is that local meaning is established prior to global meaning, as has been stated explicitly from time to time (Kintsch & van Dijk, 1978; Mitchell & Green, 1978). This follows from the view that the meanings of phrases are composed of the meanings of the individual morphemes making up the phrases, and that sentence meaning is composed from those meanings.

A third commonplace assumption is that interpretation proceeds in a left-right, incremental fashion. The strongest claim of this sort is that made by Carpenter and Just (1983) in their immediacy assumption: each word is supposedly analyzed as deeply as possible on being encountered (fixated by the eye). There certainly is evidence which suggests that word-by-word analysis is plausible. For instance, studies of eye movements during reading (Frazier & Rayner, 1990; Rayner & Frazier, 1989) suggest the influence of preceding context in selecting the appropriate sense of a lexically ambiguous word. Reading time differences can be obtained on the actual word in question (depending on prior material), a finding consistent with some kind of immediate incremental analysis.

Of course, the idea of word-by-word analysis and integration is consistent with the idea that local meanings are built up from the meanings of the individual words, which is consistent with the compositionality principle. So, returning to our example of *Mary ate some rocks for dinner*, on the immediacy assumption, the anomalous filler *rocks* should be detected as anomalous shortly after it is encountered, since it will be interpreted as deeply as possible and will be tested against the selection restrictions on the slot representation to establish local coherence. That the sentence does sound so odd is most readily understood in terms of the operation of these three conventional constraints.

This research was initiated under ESRC Grant R000231492 to A.J.S. and was subsequently supported by an ESRC studentship to S.B.B. Thanks are due Linda Moxey, Simon Garrod, Hans Kamp, and Alan Kennedy, and three anonymous reviewers, for valuable comments on earlier versions of the paper. Correspondence should be addressed to A. J. Sanford, Department of Psychology, University of Glasgow, Glasgow G12 8RT, Scotland.

All of these assumptions are currently under question, and the present studies suggest that they are all incorrect in a serious way. Our starting point is that while it is easy to produce examples in which anomalies are readily detectable, the fact is that there are cases in which anomalies are *not* detected, and in which processing is somehow shallow or incomplete. Most of the existing evidence for such shallow processing comes from problems in which the task is to explicitly answer a question or, more commonly, verify a statement through the use of general knowledge. Erickson and Mattson (1981) described the Moses illusion, in which people tend to miss the error when answering the question *How many animals of each kind did Moses put on the Ark*? (None: it was Noah). The probability of detection was found to depend on the semantic similarity between the target word (*Moses*) and the correct item in long-term memory (*Noah*) (Erickson & Mattson, 1981; van Oostendorp & de Mul, 1990; van Oostendorp & Kok, 1990). The higher the similarity, the lower the detection rate. Effects of the Moses illusion type may be obtained with a wide range of materials, and they are not due to subjects' withholding responses on the grounds that the "errors" may have been unintentionally introduced by the writer of the materials (Reder & Kusbit, 1991). Even when subjects are instructed to watch out for anomalies, detection failures occur with a high frequency.

These errors may be linked to other observations about sentence verification. For instance, Anderson (1983) showed that if subjects learned arbitrary propositions such as *The cat attacked the snake*, and subsequently were given a speeded judgment task in which they had to verify sentences such as *A cat is a snake*, they would often erroneously classify the latter as true. This Anderson ascribed to subjects' not carrying out a full semantic analysis of the statement for verification if the concepts out of which the statement was formed were highly related. Earlier work on the verification of general knowledge statements led Smith, Shoben, and Rips (1974) to propose that under some circumstances, subjects would make a judgment on the basis of the semantic overlap of the concepts being evaluated. So there is a tendency to erroneously answer affirmatively to propositions such as *A whale is a fish* (see also Reder, 1982, 1987). The force of these arguments is that with verification tasks, subjects may only partially compute the semantic details of the concepts being compared. If there is a good match on the basis of a few features, further processing may not necessarily take place.

The Moses illusion has been used as support for two views of text comprehension. Generalizing from the question-answering data, van Oostendorp and den Uyl (1984) suggested that during reading, subjects monitor the conceptual cohesion of the mental representation of the discourse (this they define as the number and strength of the relations between the concepts or facts involved in the representation). They further suggest that the initial evaluation of a representation depends on the strength and number of connections in working memory (what they term *conceptual cohesion*) and not on the specific nature of the connections (*semantic coherence*). A similar distinction has been made by Sanford and Garrod (1989) for mechanisms of anaphoric reference resolution. If the global fit of concepts in working memory (conceptual cohesion) is high, then more detailed, effortful, time-consuming analysis may not take place. Thus global goodness of fit is used to guide depth of processing, according to this model. A similar stance is taken by Reder (1987). Obviously, the Moses illusion is compatible with this view: The term *Moses* provides a good global fit to the memory representation (through high featural similarity, it is claimed), and information associated with Moses is accordingly not analyzed in sufficient detail for the anomaly to be reliably detected.

The PDP sentence-processing model described by McClelland, St. John, and Taraban (1989) provides a natural framework for such partial processing phenomena, relying as it does on multiple soft-constraint satisfaction. A major feature of this model is that the contextual influence of the current representation of a sentence is assumed to determine the extent to which a newly encountered word could have an impact on this current representation. With reference to the Moses illusion, they suggest that the constraints associated with the word Moses are not sufficiently strong to override the constraints imposed by the context. We shall not describe the model any further here, but rather note two important points that McClelland et al. make. First, they claim that the effect is incompatible with a strong notion of compositionality, in which the meanings of each word are retrieved and combined to obtain the meaning of the sentence. Rather, the contributions of words to sentence meaning is seen as graded. Second, the model explicitly breaks the distinction between word meaning and more general aspects of significance, thus departing from the view that local meaning establishment precedes more global meaning establishment (significance). In McClelland et al.'s model, global meaning may influence the contribution of a local element to the discourse, such as the meaning of a newly encountered word. The model is therefore an example of how the claim that global goodness of fit controls the degree of analysis afforded new input might be implemented, although the model is entirely passive, of course. Although in the specific model that they describe, words begin to have an effect as they are encountered, the model is not strictly left-right, since to be strictly left-right there has to be some notion of complete processing of one word before the next is encountered, and this idea has no place in the model.

In the present paper, we examine the claim that global goodness of fit influences the extent of subsequent analysis in much more detail. Rather than use a test of statements against memory, we explore factors influencing the likelihood of subjects detecting an anomalous case role filler spontaneously in a piece of discourse under straightforward reading conditions. Our studies are manipulations of the following text:

> There was a tourist flight travelling from Vienna to Barcelona. On the last leg of the journey, it developed engine trouble. Over the Pyrenees, the pilot started to lose control. The plane eventually crashed right on the border. Wreckage was equally strewn in France and Spain. The authorities were trying to decide where to bury the survivors. (1)

The fact that many people fail to notice that the query is about burying survivors shows that failures to detect anomalies do occur during normal reading rather than in testing assertions against facts stored in memory. This assumption is made by investigators who believe that the Moses illusion says something about normal discourse processing, but it is never explicitly investigated.

Our first objective in these experiments was to test the idea that the extent of semantic analysis that the critical item receives is a function of the general fit of the item to expectation based on context. In Experiment 1, we tested the idea that if the semantics of the target words (the anomaly) make the information *is alive* readily available, anomaly detection will be more likely than it will if this feature was not readily available but merely presupposed. The test was carried out both by manipulating the critical noun involved, and by using adjectival qualifications. Experiment 1 can be thought of as analogous to the semantic similarity tests carried out in relation to the Moses illusion (e.g., the sentence verification work of van Oostendorp & de Mul, 1990). However, the role of adjectival qualification was examined for the first time in the present study. Should global goodness of fit reduce further analysis, it might be possible to restrict the detailed analysis of all elements of a noun phrase as a whole if there was a fragment of the noun phrase that produced a good fit. In Experiment 2, we tested this possibility by using the expression *surviving dead*. Since *dead* provides a perfect fit to expectation, we conjectured that this degree of fit would suppress further analysis of the noun phrase, so that it would not be detected as internally anomalous. It provided a strong test of the the idea that global good fit can produce a reduction in further analysis. This manipulation had other implications too. If a local anomaly such as *surviving dead* has a low detection rate, this is good evidence that local meaning (noun phrase level) does not have to be established prior to the incorporation of the noun phrase into the rest of the discourse. It also has implications for the immediacy assumption.

Whereas Experiments 1 and 2 were based on the assumption that the relative availability of an *is alive* component and (in Experiment 2) an *is dead* component underlie anomaly detection rate, in Experiment 3 we tested the idea that the high relevance of the word *survivors* to the aircrash would produce the good fit. The scenario was changed to manipulate the relevance of *survivors*. Furthermore, we manipulated the order in which information leading to expectations occurred. Finally, if a good global fit may reduce more detailed semantic analysis, is this true of other sources of coherence? If coherence is high for one reason, are other aspects of the text that support coherence somehow inhibited? This more general question was investigated in Experiment 4.

The experiments as a whole were intended to provide more detail on the way in which partial processing occurs, and to add to our understanding of what conditions cause processing to be shallow or deep. Apart from the relatively crude manipulation of linguistic focus carried out by Bredart and Modolo (1988), there has been relatively little exploration of the properties of discourse that influence degree of processing, although much has been made of the basic Moses illusion as a rebuttal of both the completeness and local-to-global processing assumptions described earlier. Our studies depart from those in the existing literature in that we employed a single case method because we did not want subjects to use a special strategy of consciously checking for anomalies. Performance is known to be influenced when multiple items are presented (Reder & Kusbit, 1991). Our technique consisted of a single shot, with debriefing. For all of the manipulations, this was essential, since we needed to know precisely what the subjects had noticed, and how they actually interpreted the passage.

EXPERIMENT 1

The term *survivor*, although a salient word in the context of an aircrash, has a dictionary meaning of one who lives beyond some event: being alive is part of the definition of the term. Other related terms, such as *injured*, have meanings based on one's being damaged in some way, but being alive is not part of the definition. However, when speakers use the terms *the survivors* or *the injured*, a listener would presuppose that they were used of people who were in fact alive at the point when they were refered to. It seems likely, therefore, that the feature *is alive* would be accessed more readily when *survivor* is encountered than is the presupposition *is alive* when the term *injured* is encountered. If this is the case, anomaly detection rates should be better for *survivors* than for *injured*. A pretest was carried out to establish the empirical validity of the meaning-presupposition distinction drawn above. Then, in the first manipulation, comparisons of detection rate for the term *survivors* were made with three other terms: *injured*, *wounded*, and *maimed* (called the *injured* group). In the second manipulation, *injured*, *maimed*, and *wounded* were qualified by the adjective *surviving*. The prediction was that the adjective should increase the likelihood of detection, because it carries the feature *is alive*.

Pretest

The experiment depended on the term *survivors'* explicitly having a highly available *is alive* feature, whereas the *injured* terms only carried the presupposition that they were used of a person who was alive. This difference should emerge in the definitions of these terms produced by subjects. A short questionnaire was constructed. On the first page, the subjects were asked to define the terms *injured*, *wounded*, *maimed*, and *survivor* with respect to an aircrash (asked as four separate questions). It was anticipated that *is alive* would be explicitly mentioned as part of the meaning of *survivor* only, and that it would not be mentioned with reference to the other terms. Eighteen subjects did the task, undergraduate students who did not take part in any of the main experiments. The results, shown in Table 1, indicate that the *injured* group of expressions did not cause *is alive* to explicitly come to mind, but that this feature was an explicit part of the meaning of *survivor*. In giving definitions for the *injured* group, the subjects mentioned kinds of damage that the victims had suffered. A second sheet, seen subsequently by the same subjects, asked these questions: If there was an aircrash involving many people, and a news reporter used the term *the survivors/injured/maimed/wounded*, do you think that she would be using the term of people who were definitely alive at the time she used it? (yes/no). The four options were presented as four separate questions in random order. This was done to examine whether being alive would be *presupposed* when one of these words was used. The results, in Table 1, show that all subjects checked "yes." So, the *injured* terms *presupposed* that the persons described by these terms were alive, but *being alive* was not a central part of the meaning of the terms. In contrast, *being alive* was a central part of the meaning of the word *survivors*.

Method

Materials and Procedure. In each case, the material used was the basic passage shown in Text 1, followed by the question "What should the authorities do?" Four conditions were produced through the substitution of *injured, maimed*, or *wounded* for *survivors*. These four constitute the simple noun phrase (simple NP) set. Three more versions were constructed in which the critical items were *surviving injured, surviving wounded*, and *surviving maimed*; these constitute the adjectivally qualified set (qual NP). A given subject saw only one version of the problem, the whole design being independent groups.

A given version was presented on a Macintosh computer adapted for self-paced reading. By pressing the space bar, the subjects could pace the rate at which they saw successive sentences. At the end of the text, the question was preceded by the warning **QUESTION**. The subjects were told to write their solution to the problem down on a piece of paper when they encountered the question. Prior to seeing the main passage, the subjects had experience of reading a different passage according to the same procedure. There was no anomaly in the first passage, and no indication was given that the second passage would contain an anomaly. After completing the task of writing down a solution, the subjects were given a debriefing interview. They were first asked whether they had noticed anything odd about the passage. If they had not, it was put to them that the word *survivors* (etc.) had been used. They were asked whether they had noticed this. If they had not, they were asked whether they thought this made the passage odd. At debriefing, the subjects were asked whether they had ever encountered this, or any similar problem, before. Those who had encountered any version of the survivors problem were excluded from the analysis.

Subjects. One hundred fifty subjects were tested. They were mostly undergraduates, with some postgraduates from the University of Glasgow. The majority were enlisted from first-year classes in general psychology. All the subjects were naive as to the aims of the experiment, and to psycholinguistics in general.

Table 1
Pretest Results for Experiment 1

Term	Test 1: Subjects Mentioning "Is Alive"	Test 2: Subjects Presupposing "Is Alive"
Survivor	18 (max)	18 (max)
Injured	3	18
Wounded	2	18
Maimed	1	18

Results and Discussion

The subjects underwent a structured debriefing, which was used in conjunction with the written responses as the basis for making the detect/nondetect distinction. If a subject's protocol explicitly mentioned the anomaly, the subject was classed as a detector. If not, the subject was asked if he/she had noticed anything strange about the wording of the passage. If so, but if the subject had produced a written response which did not reveal this, he/she was asked why. This occurred on a very small number of occasions; it was designated as constituting *cooperative responses*, and such respondents were also classified as detectors. If a subject did not comment on the anomaly, that bit of the wording was pointed out. The subject was asked whether he/she had noticed this. All subjects reaching this stage in debriefing expressed surprise, typically commenting that they had missed the anomaly completely. These subjects were classed as nondetectors. Also, all subjects were asked if they had come across the joke about burying the survivors before, in any form. The subjects were confident about whether they had or had not, and those who had seen it before were eliminated from the analysis. In all, 47 subjects claimed to have seen some version of the problem before and were eliminated from the analysis.

Finally, no person would have been classified as producing a detection failure if he/she said that he/she thought *the survivors, injured*, and so forth, had "died later"; however, such interpretations never occurred either in the written responses or in the debriefing in this or any other experiment.

Written solution content. Most of the detectors' solutions asserted that people who are not dead should not be buried. A few detectors did write down solutions that the dead should be buried but were picked up as detectors by the debriefing procedure; the responses were revealed as cooperative ones, in that the subject had assumed an error in the presentation of the materials. The nondetectors' solutions involved relatives' decisions or home towns as part of the reasoning. A sample of solutions is presented in Table 2.

Detection rates. Figure 1 shows the overall detection rates. Consider first the simple NP group: the overall detection rate of 30% shows that the anomalies were easily missed. The term *survivors* produced a higher detection rate than did the other expressions, in accord with the hy-

Table 2
Sample of Written Solutions From Experiments 1 and 2, With Condition Indicated

Condition	Written Solutions
	Detectors
Survivor	Survivors would not need to be buried.
Injured	I wouldn't bury the survivors because they're not dead.
Wounded	The wounded would not need to be buried as they are obviously not dead.
Surviving injured	You don't bury surviving injured people.
Surviving maimed	The surviving maimed wouldn't need to be buried.
Surviving dead	You don't get surviving dead. Don't bury the surviving dead since they're not dead.
	Nondetectors
Survivor	Ask the relatives of the deceased where they would prefer them to be buried.
Injured	Find the country of origin of the dead, and bury them in whatever country is closest to their homeland.
Wounded	Let the relatives decide where they want the bodies to be buried.
Surviving maimed	Ship all the bodies home.
Surviving wounded	Contact the next of kin of all the dead.
Surviving dead	Bury the dead back in their own country. Have the relatives decide.

pothesis, and a chi-square test showed an effect of conditions on detection [$\chi^2(3) = 11.09, p < .011$]. The effect is attributable to the difference between the *survivors* and the other three (the *injured* group). The three detection rates of the *injured* group did not differ reliably from each other [$\chi^2(2) = 1.8$, n.s.], but that for *survivors* differed from these when they were pooled [$\chi^2(2) = 7.86, p < .005$].

Turning to the three qual NP examples, there were no differences in detection rates among these [$\chi^2(2) = 1.19$, n.s.]. When pooled, this set had an average detection rate of 66%, higher than the average for the three unqualified counterparts (injured, maimed, wounded) at 17.5%. The difference between the two pooled sets is reliable [$\chi^2(2) = 20.04, p < .01$].

Thus the two main hypotheses were confirmed. First, detection rates were lower for the *injured* terms than for the term *survivors*, and second, by putting in an adjectival qualification (*surviving*), detection levels were brought back up to the level for *survivors* itself. These findings are consonant with the idea that when the critical items are encountered, they receive an analysis that is often superficial. Although the term *survivors* has *is alive* as a central part of its meaning, it still produces detection failures. Presumably, the mere fact that *survivors* and *dead* are associated at all is a sufficient degree of relatedness to satisfy the crude level of analysis underlying the establishment of cohesion. When *is alive* was merely presupposed, as in the simple NP *injured* group, detection rates were even lower, as predicted. This finding confirms the view that if the *is alive* information is less readily accessed, it is less likely to be entered into the matching process. The higher level of detection for the qual NP items fits this picture.

EXPERIMENT 2

The partial matching account of the process assumes that the processor will rely on a minimum amount of semantic overlap between the critical term and the role specification in order to accept the term as a role filler. Only in the presence of available contradictory evidence is a poor fit registered. This implies that further semantic analysis of the critical item does not take place after a fit has been accepted; the details of the meaning of *survivor* are not accessed at all, otherwise detection would take place. If this is the case for a single word, it may also be the case for a qualified noun phrase. That is, if there is a good fit between a noun in a qualified noun phrase and the role slot, further analysis of an adjective might not take place. In the light of this argument, consider the phrase *surviving dead*. This is anomalous in its own right, and according to an account of processing in which it is assumed that the local semantics of a nounphrase are computed prior to its incorporation into a more global text representation, the anomaly would be detected at the early stage of processing. On the other hand, all of the data up to now suggest that for single words local processing is incomplete, and dependent on strong contextual constraints. If we think of *surviving dead* as a bundle of semantic features, then *dead* offers a perfect fit to the role slot in our scenario. It is possible, therefore, that using *surviving dead* as a critical item would produce much reduced detection rates. Of course, the strength of this argument depends upon the relative speed with which the *is alive* component of *surviving* becomes available, and the completion of establishing the good fit between *dead* and the role slot.

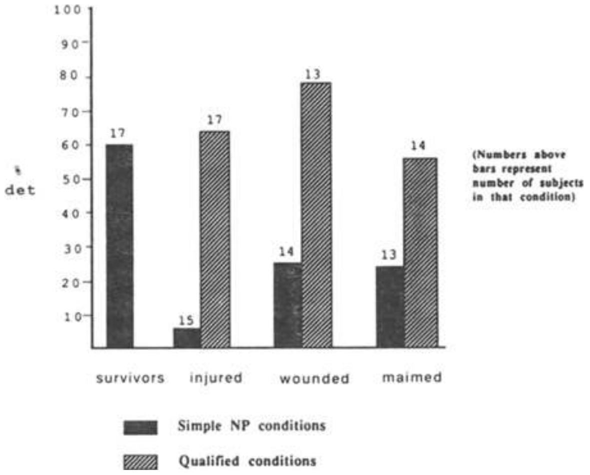

Figure 1. Detection rates from Experiment 1.

Experiment 1 showed that the element *surviving* only guaranteed about 66% detection, and that it was therefore an imperfect cue against accepting the fit of the phrase in which it occurred to the role slot. In contrast, the word *dead* is not anomalous at all: it is a perfect fit. There are therefore reasonable a priori grounds for supposing that *surviving dead* would yield poor detection rates, if role-filler processing was incomplete. Such a result would be inconsistent with the view that local semantic coherence is established prior to incorporation into the global text representation.

The term *surviving dead* was investigated alone. It can be thought of as an extra condition to Experiment 1, which was carried out concurrently. The expression is in itself judged to be anomalous. Thus 12 respondents, students of psychology, were asked if there was anything wrong with this noun phrase by itself, and all said that it was contradictory or obviously unacceptable.

Method

Material and Procedure. The procedure was identical to that used in Experiment 1. The basic passage was presented to each subject, with the term *surviving dead* substituted for *survivors*. Debriefing followed completion of the test question.

Subjects. A total of 40 subjects were tested from the same pool as in the previous experiment. At debriefing, 5 were discovered to have encountered some version of the survivors problem before. They were eliminated from the present analysis.

Results and Discussion

The written responses were similar to those obtained in Experiment 1 (a small but representative sample is shown in Table 2). The debriefing procedure was the same as that in Experiment 1. A particular concern in this study was the possibility that subjects might interpret *surviving dead* as meaning *intact dead bodies*, or something similar. No subject at all put this interpretation on the passage. Nondetectors simply reported having "missed" the anomaly. The detection rate was only 23%, numerically less even than the baseline *survivors* example. A comparison of the *surviving dead* condition with the pooled adjectivally qualified examples from Experiment 1 (which gave 66% detection), produced a reliable difference [$\chi^2(1) = 55.1, p < .001$].

The extremely low detection rate suggests that the local semantics of the phrase *surviving dead* are not computed prior to their incorporation into the more global representation of the text. If they were, the anomaly should be noted at that initial stage. In the case of the qual NP items in Experiment 1, the adjective had the effect of enhancing anomaly detection. In the present case, it had no such effect. This is as expected on the basis that *dead* provides such a good fit that further analysis of the noun phrase is suppressed. Our argument is that just as the good global fit of *injured*, *survivors*, and so forth suppressed further analysis at the single word-meaning level, so the good fit of *dead* in the present case suppressed the analysis of its qualifier. Thus depth of analysis effects carry over to qualified noun phrases.

EXPERIMENT 3

The discussion up to now has been based on the idea that expectations regarding the presence of dead victims influence the impact of the critical items: the poorer the fit of a critical item to some representation of dead victims, the better the detection rate. This experiment was a direct manipulation of expectation by varying the underlying scenario, so as to influence the plausibility of accident victims' being dead.

Because of the large number of conditions to be investigated, a modified procedure was used. Rather than a full narrative text, the critical question was embedded in a questionnaire of 10 questions about social issues. All were one-line questions, such as "At what age should it be legal for people to buy cigarettes?" The subjects were encouraged to write what they thought about such issues. The fifth question was the critical one, containing the anomaly (e.g., "When an aircraft crashes, where should the survivors be buried?"), and it could have the scenario information presented prior or subsequent to the anomalous *survivors*. This expression could also be embedded within an active ("...bury the survivors") or passive ("...survivors be buried") verb phrase, which provides a manipulation of the relative position of the verb. Hence, there were four question types:

1. Early scenario, passive VP (late verb)

When (an aircraft crashes/a bicycle accident occurs), where should the survivors be buried?

2. Early scenario, active VP (early verb)

When (an aircraft crashes/a bicycle accident occurs), where should you bury the survivors?

3. Late scenario, passive VP (late verb)

Where should the survivors be buried after (an aircrash/a bicycle accident)?

4. Late scenario, active VP (early verb)

Where should you bury the survivors of (an aircrash/a bicycle accident)?

The global effect of scenario, independent of order, and so of any left-right processing assumption, can be tested by considering the overall rates for the scenario manipulation across question types. By analyzing question types, one can test for any effects of the order in which constraining information is introduced. A strictly left-right processing account would predict that the greater the contextual information *prior* to the critical word, the smaller should be the impact of the anomalous expression, and hence the lower the likelihood of detection. Under such an incremental view of interpretation, detection rate should have been highest in Condition 3, in which *survivors* preceded both the scenario and the verb, and it should have been lowest in Condition 2, in which it was preceded by the others. The rates for Conditions 1 and 4 should have depended on the relative impact of the verb and the scenario.

Pretest

The logic of the design depended on the presence of dead victims being much more likely in the case of an aircrash than in the case of a bicycle crash. The validity of this assumption was tested by asking 16 subjects to respond to a questionnaire that probed their perception of the likelihood of dead victims resulting from five different kinds of accidents. The critical pair for the present purpose was *a passenger carrying airplane crash* and *a bicycle accident*. The subjects were asked to select the most appropriate answer to each question against five possibilities (certain/probable/possible/unlikely/definitely not). A clear difference was obtained: for the aircrash, 14 said "certain" and 2 reported "probable." For the bicycle crash, 11 reported "possible" and 5 reported "unlikely." Thus, whereas dead victims are considered possible or unlikely following a bicycle accident, they are not considered probable. If one treats the statements checked as ordered in terms of certainty, there was no overlap between response categories for the two scenarios (sign test, $p < .01$ for $N = 16$, $x = 0$). The scenarios differed in the way expected.

Method

Materials and Procedure. The subjects were asked to fill out a 10-item questionnaire seeking their views on social and related issues. The 5th question was always the one containing the anomaly. After each subject had filled in the answers, he/she was debriefed in the standard fashion. Testing continued until a total of 15 subjects, without previous knowledge of the problem, had been tested in each cell of the design (i.e., a total of 120).

Subjects. These were undergraduates at the University of Glasgow. They had no knowledge of the purpose of the study or of psycholinguistics in general.

Results and Discussion

The detection rates are shown in Table 3. First, it is evident that poor detection rates obtain in the questionnaire version with aircrash scenario, just as they did with the fuller description used in Experiments 1 and 2. A 2 (scenario type) × 2 (scenario: early/late) × 2 (verb: passive/active) analysis of variance based on the partitioning of chi-square was carried out on the detection rate data (Winer, 1971).

Despite numerical trends in the data, the only reliable result was that of scenario type. Detection rate was considerably lower in the aircrash case than in the bicycle case, as expected [$\chi^2(1) = 11.53$, $p < .01$]. No other main effect or interaction approached significance (all χ^2s < 2).

Since no effect was found for any order manipulation, there is no evidence to support strictly incremental interpretation. In order to provide a better test, a further 133 subjects were tested across the four conditions, but with the aircrash scenario only. These results, combined with results from the main study, are shown in Table 3. The overall level of detection at 37% compares well with the original overall rate at 33%. There is no sign that presenting the scenario first resulted in a lower detection rate [overall $\chi^2(3) = 5.49$, n.s.]; on the contrary, there does appear to to have been a trend toward higher detection rate with the active constructions. A post hoc test of this alone gave $\chi^2(1) = 4.41$, $p < .05$. A similar post hoc partitioning of the early-late scenario factor yielded an insignificant result. Although there is marginal evidence for a passive/active difference, the results show a higher detection rate when the verb preceded the critical item (active construction), which is good evidence against the proposed order effect. The marginal effect was only just reliable on the post hoc test, and further investigation would be necessary to make any stronger claims on the basis of it.

Taken together, the results show a strong global effect of scenario type, which conforms to the idea that detectability is a function of scenario-based expectation. However, detection rate is not lower if the scenario is introduced prior to the critical item, or if the verb is introduced prior, or both. This suggests that the critical item is not *fully* analyzed as it is encountered, before one moves on to subsequent material, but that later input (scenario information) can influence the development of the impact of the critical item.

Table 3
Detection Rates for Experiment 3 (in Percent)

	Scenario	
Question Type	Aircrash	Bicyclecrash
Early scenario, passive VP	20 (26)	80
Early scenario, active VP	27 (44)	73
Late scenario, passive VP	33 (31)	73
Late scenario, active VP	53 (46)	93
Scenario average	33 (37)	80

Note—Aircrash figures in parentheses indicate result of enlarging the sample.

EXPERIMENT 4

In this experiment, we return to the claim that once the system has a satisfactory level of information supporting coherence, further analysis might not necessarily take place. Specifically, we tested the idea that if information relevant to the question of place of burial was easily accessed, then a deep analysis of the term *survivors* would not take place. Only if a more thorough search for information relevant to the question was required would a deep analysis of the term be likely. What constitutes relevant information can be gleaned from an analysis of the response protocols in the experiments up to now. These show the nondetection responses to be answers that took the victim's country or town of domicile into account, and information having a bearing on this was used in the high-relevance condition of the experiment. In order to provide the extra information, a two-sentence version of the questionnaire format was employed, the basic version of which was as follows:

> Suppose that there was an aircrash with many survivors.
> Where should they be buried? (2)

This format allows the description of the situation, or of the survivors, to be elaborated by the simple extension to relative clauses. An example situational elaboration is the following:

> Suppose that there was an aircrash with survivors, *which happened last week*. Where should they be buried? (3)

In this case, the additional information is not very helpful in answering the question. However, in the following case it is:

> Suppose that there was an aircrash with survivors who were mostly European. Where should they be buried? (4)

A manipulation such as that in Text 4 enables us to test the conjecture that providing information *relevant to answering the question* will cause this information to dominate processing, and according to our hypothesis, the impact of the semantics of survivor will be lessened. This would manifest itself as lower spontaneous detection rates when relevant information is provided.

Method

Conditions. All critical items were presented in the two-sentence format described above. The conditions are described in Table 4, along with a brief justification. Conditions 1 and 2 will provide baseline information; Conditions 3 and 4 provide irrelevant qualification information. If information irrelevant to answering the question has no effect on detection rate, then Conditions 1, 2, 3, and 4 should have equivalent rates, and can then be treated as a general baseline. Conditions 5-7 provide a test of the relevance argument, and it is expected that these will give lower detection rates than baseline. Condition 8 provides a control in which the qualification indicates explicitly that the burial question is not appropriate, and so should produce a higher detection rate than should the other conditions. In this condition, the *is alive* component is reinforced by a comment on the state of health of the survivors.

Table 4
Conditions in Experiment 4

A. Basic versions:
1. . . . *survivors*.
2. . . . *many survivors* (less stilted than *survivors* alone, but the referent of *they* must still be those who survived).

B. Question-irrelevant qualifications:
3. . . . *survivors who were mostly gravediggers* (a qualification on the survivors, but of only marginal potential as a piece of relevant information).
4. . . . *survivors, which happened last week* (irrelevant qualification for the question).

C. Question-relevant qualifications:
5. . . . *survivors who were mostly European* (relevant qualification for the question).
6. . . . *survivors who were mostly of no fixed abode* (relevant qualification for the question).
7. . . . *survivors who were mostly circus performers* (potentially relevant qualification, since circus performers are typically itinerant, but requiring this inference to be made).

D. Anomaly-focussing control:
8. . . . *survivors who were mostly unhurt* (a qualification which makes the burial question inappropriate).

Procedure. The material for a given condition was embedded in a questionnaire, as for Experiment 3, but the questions were all modified so as to be in the two-sentence format. Only one version of the critical question appeared in a given questionnaire, and a given subject received only one questionnaire. The subjects were given the same instructions as for Experiment 3.

Subjects. These were mostly students at the University of Glasgow. They were unaware of the purpose of the investigation. Testing was continued until there were at least 15 subjects in each condition who had not previously encountered the problem. A total of 205 took part.

Results and Discussion

The proportion of anomaly detection under each condition is shown in Figure 2. It is apparent that when information relevant to answering the question was given, there was a drop in detection rate. Overall, there was a strong influence of condition on detection rate [$\chi^2(7) = 16.94, p < .05$]. A variety of separate partitionings were undertaken along the lines decided in advance. First, Conditions 1-4 (no qualification, or irrelevant qualification) were compared, and these were not reliably different [$\chi^2(3) = 1.13$, n.s.]. Detection rate is not measurably altered by *simply* qualifying either the event or the survivors, and this group can therefore be considered a baseline, with an overall detection rate of 76%. Second, Conditions 5-7, all of which included additional relevant information, did not measurably differ [$\chi^2(2) = 0.38$, n.s.). The reduced level of detection for this group as a whole (49%) was compared with the pooled baseline data. This difference is reliable [$\chi^2(1) = 8.64, p < .01$]. Finally, the result of Condition 8 shows that qualifying information which requires individuals to be alive produced a detection rate of 100%, which is reliably better than both the relevant conditions [$\chi^2(1) = 12.5, p < .001$] and the baseline conditions [$\chi^2(1) = 4.53, p < .05$].

The findings are generally supportive of the relevance hypothesis: if information pertinent to answering the question is made easily available, this provides a level of coherence satisfactory to the comprehension system, and the critical term *survivors* receives only a cursory analysis. Detection rates are thus low. If such information is not readily available, possible sources of information (including word meanings) are explored in more detail by the system, and detection rates are higher. If the more detailed analysis leads to information supporting the *is alive* component (Condition 8), detection rates are higher still.

GENERAL DISCUSSION

The strength of the anomaly detection paradigm is that it reveals underlying shallow processing. It has already been argued that failures to detect anomalies such as these is consistent with the view that in normal comprehension, exhaustive tests of the fit of a filler to a role are not carried out; indeed, it has been claimed that exhaustive checking of attributes is neither reasonable nor feasible (Erickson & Mattson, 1981; McClelland et al., 1989). Other work (Reder & Kusbit, 1991), and the procedures em-

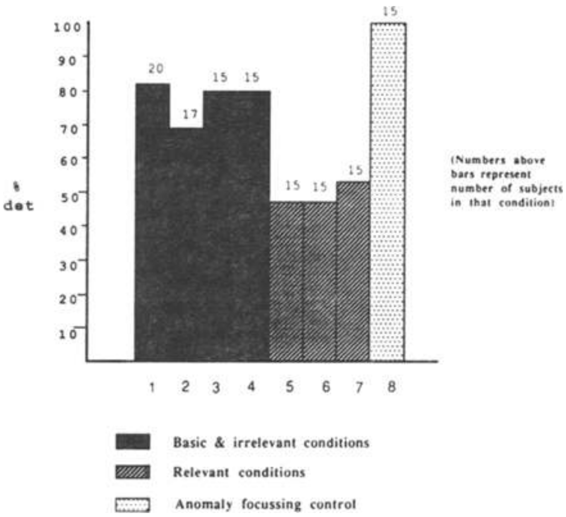

Figure 2: Detection rates from Experiment 4.

ployed in the present case study, show the failures to notice to be genuine, and not simply failures to report, based on an assumption by the reader of unintended writer error. Rather, we suppose that the comprehension system will accept partial matches, or partial analyses, assuming coherence as a default. To the extent that an anomaly is not detected, we assume that the processing of its semantics has been relatively shallow.

The main issue of interest was to determine the factors controlling the degree of processing afforded to the critical role fillers in the aircrash passage, and to compare these with similar (but different) factors in work on the Moses illusion. In Experiments 1 and 2, item-based influences were observed that fit with previous work. Items that are globally consistent with statements about the victims of an aircrash (*survivors; wounded*) are sometimes missed as anomalies. If a superficial filler/role fit is obtained, perhaps in terms of feature overlap, the specifics of the semantic relation that the filler bears to the role may not be calculated, in a way analogous to that found for question-answering by Smith et al. (1974). In Experiment 1, some limitations of this were tested. It was found that if the critical item has *is alive* as a key part of its meaning (*survivors*), there is an increased probability of detecting that item in comparison with one in which being alive is merely presupposed (*injured, maimed, wounded*). These results show that the balance between a good global fit and highly available information suggesting that the item is anomalous is one factor in determining the likelihood of detection. With the adjectivally qualified conditions in Experiment 1, there was still a failure to guarantee detection. This shows that detection failure is not limited to single words, but that whole noun phrases may be analyzed in a shallow manner too. This was investigated in detail in Experiment 2, in which it was discovered that the internally anomalous noun phrase *surviving dead* led to very low levels of detection. This result was predicted on the grounds of the perfect fit of *dead* at a global level, and the preestablished fact that the adjective *surviving* did not guarantee detection: together, the balance should favor nondetection.

This result has a further implication. Many accounts of interpretation assume that local meaning is established prior to global meaning, as discussed. In such a case, it would be expected that *surviving dead* would be readily recognized as anomalous in its own right. However, our findings indicate that the local meaning analysis is inhibited in some way by the good fit of *dead* at the global level. We are not suggesting that *no* local processing occurs, only that it is effectively swamped by the good fit of *dead* in the present case. It is fairly obvious, however, that the present result is grossly inconsistent with any *obligatory* bottom-up semantic composition process. Although the result with *surviving dead* seems very surprising at first glance, we have every confidence in it, since it has been replicated using audio presentation (Barton & Sanford, 1993).

All of these arguments depend on a high expectation of dead people's being part of the scenario evoked by the aircrash topic. This makes the use of the term *survivors* pertinent. In Experiment 3, it was shown that dead people are expected in aircrashes, but that death is much less predicted in the case of a bicycle crash, and that this pattern influences detection rates. So a default expectation of many dead people, and hence the corresponding relevance of the term *survivors*, reduces the likelihood of spotting the anomaly. In addition, we tested the idea that if the aircrash scenario (and/or the verb *bury*) was introduced prior to the critical item *survivors*, then the anomaly would be less detectable because of prior expectation. This would be expected on any account in which it was assumed that incremental interpretation (left-right analysis) took place. The results obtained are inconsistent with such a view: not only was there no advantage of having the scenario and/or the verb first, there was a marginal advantage in detection for the reverse order.

So, in contrast to strict incremental interpretation, scenario effects can influence the processing afforded to immediately preceding material. These findings are consistent with a view that the amount or type of analysis required of a word (its contribution to meaning) cannot generally be known before a specific context for a specific aspect of interpretation has been encountered. But it does suggest that the contribution of a word will be small until that context is encountered. Thus, in the present study, the contribution of *survivors* is weak until a context for its interpretation is found (the scenario in the present case). When the context is encountered, further analysis may take place, the extent of which is determined by global goodness of fit. The results of Experiment 4 support just this view. Encoding the first sentence results in only a minimal contribution from the semantics of the term *survivors*. This contribution is not developed further on one's encountering the second sentence if other, more available and relevant coherence-supporting information, is forthcoming from elsewhere, which it is in the conditions that provide relevant additional information.

It should also be noted that the results of Experiment 2, in which *surviving dead* was used, have a bearing on order effects too. If the *surviving* component were analyzed to any extent before the term *dead* was encountered, much higher detection rates would have been expected. The claim that the *dead* component effectively reduces the contribution of *surviving* (to produce a very low detection rate) entails a deviation from strict left-right processing. These observations perhaps suggest reason for caution in the use of techniques for the on-line analysis of reading that rely heavily on assumptions of incrementality, such as word-by-word presentation methods (Just, Carpenter, & Woolley, 1982).

Taken together, all the evidence shows that coherence establishment is incomplete: when a satisfactory level of coherence has come from one source, a check on coherence based on other sources may not take place. Other findings in a rather disparate literature support a general view of this sort (Sanford & Garrod, in press; also Sanford, Barton, Moxey, & Paterson, in press). For example, as early as 1968, Schlesinger observed that in the processing of syntactically complex embedded sentences, pragmatic considerations could result in an interpretation that relied on an incomplete syntactic analysis. Wason and Reich (1979) provided experimental demonstrations that pragmatic plausibility can produce interpretations at odds with local semantic structure. Furthermore, Ehrlich and Loridant (1990) presented data showing that anomalous anaphors in the form of antonyms can also pass undetected, and Sanford and Garrod (1989) present a variety of evidence that processes underlying reference resolution may be incomplete, yet give the illusion of completeness. What is now required is a systematic exploration of the extent to which component processes underlying cohesion are completed in the service of comprehension, and that will be a major exercise.

In conclusion, the general picture is one in which processing of some terms during reading is rather shallow. If there is a good semantic match at a global level between a role filler and a role slot, further analysis need not take place. This holds, regardless of the order in which the role filler and the scenario constraining the semantics of possible role fillers is encountered. If there is a major way of achieving coherence that does not require further analysis, further analysis will likely not occur. Clearly, in the attempt to produce a process model of coherence establishment, attention must be paid to selective mechanisms, since not all potential sources of coherence are used; indeed, from a computational perspective, it might not be possible for all sources to be used, since these can be argued to form an effectively infinite set. To go further will require the development of a theory of selective processing. Existing psychological accounts, such as the minimalist theory that only inferences necessary for coherence are made, finesse the issue, since they do not define coherence (McKoon & Ratcliff, 1992; see also Sanford & Garrod, in press). Finally, although it is clear that PDP theories such as that of McClelland et al. (1989) provide a framework within which the present results may be understood, at their present stage of development, to the authors' knowledge, such theories have not been applied to selective processes in coherence establishment, though such a development would seem to be a promising direction.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Barton, S. B., & Sanford, A. J. (1993). *Incomplete processing of coherence relations with auditory presentations of text*. Manuscript submitted for publication.

Bredart, S., & Modolo, K. (1988). Moses strikes again: Focalization effects on a semantic illusion. *Acta Psychologica*, **67**, 135-144.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do when your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275-307). New York: Academic Press.

Ehrlich, M.-F., & Loridant, C. (1990, September). *Metacognitive control in the the resolution of anaphora in skilled and less-skilled comprehenders*. Paper presented at the conference of the European Society for Cognitive Psychology, Como, Italy.

Erickson, T. A., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning & Verbal Behavior*, **20**, 540-552.

Frazier, L., & Rayner, K. (1990) Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory & Language*, **29**, 181-200.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, **111**, 228-238.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, **85**, 363-394.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language & Cognitive Processes*, **4**, 287.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, **99**, 440-466.

Mitchell, D. C., & Green, D. W. (1978). The effects of context and content of immediate processing in reading. *Quarterly Journal of Experimental Psychology*, **30**, 609-637.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 779-790.

Reder, L. M. (1982). Plausibility judgements vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, **89**, 250-280.

Reder, L. M. (1987). Strategy selection in question-answering. *Cognitive Psychology*, **19**, 90-138.

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory & Language*, **30**, 385-406.

Sanford, A. J., Barton, S. B., Moxey, L. M., & Paterson, K. B. (in press). Cohesion processes, coherence, and anomaly detection. In G. Rickheit & C. Habel (Eds.), *Focus and cohesion in language comprehension*. Berlin: de Gruyter.

Sanford, A. J., & Garrod, S. C. (1989). What, when and how?: Questions of immediacy in anaphoric reference resolution. *Language & Cognitive Processes*, **4**, 235-262.

Sanford, A. J., & Garrod, S. C. (in press). Selective processing in text understanding. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics*. New York: Academic Press.

SCHLESINGER, I. M. (1968). *Sentence structure and the reading process*. The Hague: Mouton.

SMITH, E. E., SHOBEN, E. J., & RIPS, L. V. (1974). Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review*, **81**, 214-241.

VAN OOSTENDORP, H., & DE MUL, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psycholgica*, **74**, 35-46.

VAN OOSTENDORP, H., & DEN UYL, M. J. (1984, June). *Semantic relatedness effects in text processing*. Paper presented to the joint meeting of the Experimental Psychology Society and the Netherlands Psychonomic Foundation, University of Amsterdam.

VAN OOSTENDORP, H., & KOK, I. (1990). Failing to notice errors in sentences. *Language & Cognitive Processes*, **5**, 105-113.

WASON, P., & REICH, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, **31**, 591-597.

WINER, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

(Manuscript received December 27, 1991;
revision accepted for publication November 12, 1992.)