Comprehension of Deep and Surface Verbphrase Anaphors

Michael K. Tanenhaus and Greg N. Carlson

Department of Psychology, University of Rochester, New York, U.S.A.

Linguistic research on anaphora (Hankamer & Sag, 1976; Sag & Hankamer, 1984) suggests that anaphors can be divided into two classes: *Surface* anaphors that find their antecedents in some level of linguistic representation, and *deep* anaphors that find their antecedents in a discourse model or a corresponding mental representation. In three experiments, we tested the hypothesis that the syntactic form of the antecedent for a subsequent anaphor would affect the difficulty with which surface anaphors but not deep anaphors would be comprehended. In a "makes-sense" judgement task, surface anaphors were judged to make sense more often when the antecedent was introduced in a phrase that was syntactically parallel to the anaphor than when it was syntactically non-parallel. In contrast, the syntactic form of the antecedent did not affect judgements to the deep anaphors. Parallelism did, however, influence comprehension times for both types of anaphors. The results provide qualified support for the hypothesis that deep and surface anaphors access different types of representations during comprehension.

INTRODUCTION

One of the chief linguistic expressions of the general phenomenon of *anaphora* is found in the use of definite third person pronouns, such as "he" or "them". Much linguistic work on anaphora has focused on the

syntactic and semantic relations that definite pronouns can or must bear to their antecedents, and most psycholinguistic studies of anaphora have likewise focused on the comprehension of definite pronouns with antecedents in the same sentence or in prior discourse (for a review, see Garnham, 1987). As is well known, though, natural language employs a wider range of anaphoric expressions than just the definite pronouns. Some of these devices are familiar to psychologists (such as definite nounphrases with "the" as the article), whereas others are not generally known outside of linguistics (e.g. "stripping": Ross, 1967). In the end, there is a surprisingly rich array of anaphoric devices employed by English and other natural languages; many of these devices turn out to be implicitly expressed, chiefly as omitted constituents that need to be "filled in" anaphorically (as when someone says "John bought three green candies and two yellow", one understands by this two yellow *candies*, thereby "filling in" the omitted noun following the word "yellow").

Hankamer and Sag (1976) have proposed that the range of anaphoric devices can be divided into two main classes, which they label as *deep* and *surface* anaphors. Sag and Hankamer (1984) further claimed that these classes differ in terms of the level of representation that must be accessed to determine their antecedents; roughly, a deep anaphor directly accesses a conceptual level of representation in a discourse model or mental model (Garnham, 1981; Johnson-Laird, 1983; Webber, 1981), whereas surface anaphors must first access a purely linguistic level of representation. Thus, they point out, deep anaphors do not require a linguistic antecedent, because when a pronoun's interpretation is determined by non-linguistic context it is then "pragmatically controlled". "Pragmatic control" refers to situations in which an antecedent of a pronoun is presented in the context of use, but not introduced explicitly in a linguistic expression. This is common with pronouns, as when a person asks "Who is she?" at a party, gesturing to the person intended to be referred to by "she". Or, suppose two men sharing an apartment return home at the end of the day, finding the place ransacked. One can turn to the other and say "I bet the cops did it", where "it" is taken to be referring to the ransacking of the apartment, even though no words to that effect had yet been uttered.

But surface anaphors, by virtue of the kind of representation required to assign them antecedents, take a linguistically introduced antecedent, and cannot be pragmatically controlled. For instance, in the ransacking example just discussed, one of the two men cannot turn to the other and ask (felicitously), "I wonder who?" On the other hand, if the other man first comments, "Somebody ransacked our apartment!", then the question "I wonder who?" would be felicitous, as under these circumstances an appropriate antecedent ("ransacked the apartment") would have been introduced linguistically. Simply having the idea in mind of ransacking the

apartment is not sufficient to license the use of such cases of surface anaphora as these; however, as we have seen, it is sufficient to support the use of a deep anaphor.

A second difference between deep and surface anaphors is that the surface anaphors require that their linguistic antecedent be presented in a suitable form (being sensitive to aspects of linguistic form), whereas deep anaphors are not sensitive to the syntactic form of the expression introducing the antecedent, when the antecedent is introduced linguistically. This difference is one of "syntactic parallelism". Consider the phenomenon of VP-ellipsis in English, where (roughly) a repeated verbphrase may be omitted, but understood anaphorically. The antecedent itself must also be a verbphrase—a type of linguistic constituent. Consider sentence (1) from Hankamer and Sag (1976):

1. Someone had to *take the oats down to the bin*.

The italicised constituent is a verbphrase. Now, consider uttering (2) immediately after (1):

2. . . . So Sandy did ___ .

The blank indicates the position of the omitted verbphrase, which is understood in the context of (1), as taking the underscored verbphrase in (1) as its antecedent.[1] Now, suppose instead that in place of (1) we introduce the roughly synonymous (3):

3. The oats had to be *taken down to the bin*.

The (relevant) verbphrase in (3) is italicised. If sentence (3) instead precedes (2), the result is infelicitous:

3'. ??The oats had to be taken down to the bin. So Sandy did ___.

The explanation is that filling the omitted verbphrase with "taken down to the bin" results in an ill-formed structure, or at least a structure that is difficult to interpret. This indicates that VP-ellipsis is in the category of surface anaphora.

It is clear, though, that even if the message of (1) is instead encoded as (3), the concept of taking oats down to the bin remains. Thus one would

---

[1]The parallelism requirement is actually somewhat more abstract. For example, surface anaphors can be used felicitously when the tense of the verb in the antecedent differs from that of the anaphor, as in "John nearly caught a fish yesterday. Tomorrow he is sure that he will (*catch* a fish)." Another example involves "indexical" expressions, as in "John hates to take his shower before eating breakfast, but his wife actually prefers to (take *her* shower)." Sag and Hankamer (1984) discuss these and similar examples in some detail.

expect that a deep anaphor could take this concept as its antecedent. As predicted, with a deep anaphor, (4) is a coherent discourse:

4. The oats had to be taken down to the bin. So Sandy did it.

The claim that anaphors divide into deep and surface categories is particularly striking in light of the fact that, in many cases, a sentence with a deep anaphor appears to have the same interpretation as the corresponding sentence with a surface anaphor. Consider the examples in (5):

5. We wanted Mary to paint the garage.
   a. . . . but she absolutely refused.
   b. . . . but she absolutely refused to.

Both continuations, (5a) and (5b), are equivalent, yet (5a) contains a deep anaphor ("null complement anaphora"), whereas (5b) contains a surface anaphor (VP-ellipsis). Nevertheless, the claim is that the (5b) case involves recourse to a level of linguistic structure that (5a) does not.

The most extensive experimental test of the Sag–Hankamer hypothesis was conducted by Murphy (1985a). Murphy examined the effect of three variables on the comprehension of target sentences that contained closely matched deep and surface anaphors: (1) the length of the antecedent; (2) the distance between the anaphor and its antecedent; and (3) the syntactic parallelism of the antecedent. Although each manipulation influenced reading times to the target sentences, none affected the deep and surface anaphors differentially.

Murphy reasoned that the length of the antecedent should affect comprehension of surface anaphors, on the assumptions that interpreting a surface anaphor requires replacing the anaphor with a "copy" of the antecedent and copying a long phrase should take longer than copying a short phrase. Murphy also assumed that accessing a conceptual representation of the antecedent should not be affected by the length of the phrase introducing the antecedent. Thus he argued that antecedent length should not affect comprehension times to a deep anaphor if, in fact, deep anaphors are interpreted using conceptual representations.

Examples of the short and long antecedents are given in (6a) and (6b), respectively, for the surface and deep anaphors given in (6c) and (6d):

6a. Jimmy swept the tile floor behind the chair.
6b. Jimmy swept the tile floor behind the chair free of hair and cigarettes.
6c. Later, his uncle did too.
6d. Later, his uncle did it too.

There was a main effect of length, with target sentences being read faster with the short antecedents, but there was no interaction between length

and type of anaphor. Moreover, the effects of length were markedly reduced when the sentence introducing the antecedent and the sentence containing the anaphor were separated by an intervening sentence. Murphy concluded that when the verbatim form of the antecedent was available in memory, both deep and surface anaphors are interpreted by a "copying" process, whereas when the form of the antecedent is not available, "plausible reasoning" is used to construct an interpretation for the anaphor.

However, the manipulation of antecedent length also introduced potential scope and attachment ambiguities which could have complicated comprehension of both types of anaphors. Thus, the length effect alone does not provide convincing evidence for a copying process. The fact that the length effect disappears with intervening material is consistent with Murphy's hypothesis but, as we will see, the interpretation of the distance effect, itself, is problematic. We should also note that copying is not the mechanism by which a surface anaphor would be associated with its antecedent in most current linguistic theories. The more standard assumption is that anaphors are co-indexed with their antecedents, i.e. linked by pointer to the antecedent. On an indexing or pointer account, there is no reason to expect that the length of the antecedent should have different effects on deep and surface anaphors.

Murphy also varied the distance between the antecedent and the anaphor. He contrasted a "near" antecedent condition in which the sentence containing the anaphor immediately followed the sentence containing the antecedent with a "far" antecedent condition in which a sentence intervened between the anaphor and its antecedent. If surface anaphors require access to the linguistic form of their antecedents, comprehending surface anaphors in the far condition should be especially difficult, because memory for linguistic form decays rapidly, whereas memory for conceptual information is much more robust (Sachs, 1967; Wanner, 1974). Thus the amount of material that intervenes between an anaphor and its antecedent should have smaller effects on the comprehension of deep anaphors than the comprehension of surface anaphors (see Malt, 1985, and Garnham and Oakhill, 1987, for studies examining the effects of distance on interpreting surface anaphors). Murphy found that comprehension times to both surface and deep anaphors were longer in the far antecedent condition than in the near antecedent condition. Although the interaction with distance was not significant, the difference between the near and far antecedents was numerically larger for the surface anaphors than the deep anaphors.

The lack of a significant interaction between distance and type of anaphor might have been partially due to the intervening sentence having introduced material that removed the antecedent from focus. For pro-

nouns and definite nounphrase anaphora, the distance between an antecedent and its anaphor has little effect on comprehension times for the sentence containing the anaphor as long as the antecedent remains in focus (Anderson, Garrod, & Sanford, 1983; Garnham, 1987; Lesgold, Roth, & Curtis, 1979). Thus, the fact that distance effects obtained with the deep anaphors in Murphy's experiments raises the possibility that the intervening material introduced topic shifts. Tanenhaus, Carlson, and Seidenberg (1985) report an experiment in which an intervening sentence increased comprehension times to surface—but not deep—anaphors. The intervening sentences in this experiment were usually stative sentences, which as a rule do not introduce topic shifts or focus shifts. The issue of how intervening material affects the interpretation of deep and surface anaphors remains unresolved, and further research is needed to separate clearly focus and topic shifts from distance effects.

Murphy also varied the syntactic parallelism of the antecedent for deep and surface anaphors, crossing parallelism with distance and with length of the antecedent. In the near antecedent condition, both types of anaphors took longer to comprehend when their antecedents were not syntactically parallel; however, when a sentence intervened between the antecedent and the anaphor, there were no effects of parallelism.

The finding that both deep and surface anaphors are equally dependent upon the syntactic parallelism of their antecedents is clearly incompatible with any form of Sag and Hankamer's (1984) hypothesis. This result is also somewhat surprising in light of the strong linguistic intuitions that parallelism is important only for surface anaphors. In addition, the results are inconsistent with results we obtained in a study that found syntactic parallelism effects only for surface anaphors (Tanenhaus & Carlson, 1984).

The present experiments manipulated syntactic parallelism in similar ways to Murphy's experiments; however, the materials and the task were somewhat different than his. Murphy embedded his context and target sentences in short paragraphs and used a reading time paradigm. We used pairs of stimuli in which a context sentence introduced an antecedent for an anaphor in a following "target" sentence, and we used a "makes sense" judgement task in which subjects were instructed to decide as quickly as possible whether or not the target sentence made sense given the context sentence. The measures in this task are the percentage of sentences judged to make sense and latencies to those sentences that are judged to make sense. We have adopted the judgement task in recent studies (e.g. Tanenhaus & Carlson, 1984) because it requires the subject to take the antecedent into account when reading the sentence with the anaphor and because in a number of experiments in our lab more stable data is obtained with this task than with simple reading time or comprehension time tasks.

To see why this might be the case, it is important to keep in mind that

comprehension time and reading tasks are in fact variants of signal detection experiments. In the instructions to a standard comprehension time experiment, subjects are typically told to press a button when they have "understood" a sentence, leaving what is meant by "understood" up to the subject. In effect, this leaves the subject free to set his or her own criterion. We find that subjects are often confused by what we mean by "understand", and it is common lore among researchers who conduct comprehension time experiments that the type of questions that are included to make subjects pay attention during the experiment have a great deal of influence on overall comprehension times. Presumably, this is because subjects use the questions to set their criterion for comprehension; if they can answer the question, they have understood the material. The makes sense task allows the subject to set a criterion using filler sentences that do not make sense. The criterion set by the subject will determine whether differences will be reflected in judgements or latencies or both, although we have not, to date, conducted parametric studies that manipulate the subject's criterion.

We conducted three experiments investigating the effects of parallelism on the comprehension of deep and surface anaphors. The first experiment manipulated parallelism by introducing the antecedent in either an active or a passive sentence (e.g. "Someone has to take out the garbage" vs "The garbage has to be taken out"). In the active version, there is a linguistic constituent that can serve as the antecedent for the anaphor, whereas the passive version does not contain an appropriate constituent.

The second experiment adopted a different approach to manipulating parallelism. Antecedents for verbphrase anaphors were either introduced in a verbphrase or in a nounphrase using a nominalised verb. The use of a nominalised verb allows an event, which is the right type of *conceptual* object for a verbphrase anaphor, to be introduced in a nounphrase, which is the wrong type of *linguistic* category for a verbphrase anaphor. The Sag–Hankamer hypothesis predicts that only conceptual-level representation should matter for deep anaphors and thus the nominalisation manipulation should affect the surface but not the deep anaphors.

The third experiment was conducted to eliminate a confound between type of anaphor in the Hankamer–Sag taxonomy and phonological explicitness of the anaphor. In most surface anaphors, the anaphoric element is a deleted phrase and thus is not realised phonologically. In contrast, for most deep anaphors, there is an explicit anaphoric element. However, there are some types of deep anaphors in which the anaphoric element is unrealised phonologically. One such type is *null complement anaphora*. In null complement anaphora, the missing infinitive or sentential complement of some verbs can function as an anaphor, as was illustrated earlier in (5). The same verbs that allow null complement anaphora also can be used as

surface anaphors when the material following the infinitive marker ("to") is deleted. The contrast between null complement anaphora (a deep anaphor) and VP-ellipsis (a surface anaphor) was used in Experiment 3 in order to examine the effects of syntactic parallelism on deep and surface anaphors when both types of anaphors involve null elements.

EXPERIMENT 1

Parallel antecedents were introduced in active sentences and non-parallel antecedents in passive sentences. An example is given in (7):

7a. Someone had to take out the garbage.
7b. The garbage had to be taken out.
7c. But Bill refused to do it.
7d. But Bill refused to.

Sentence (7a) is the active (parallel) antecedent and (7b) is the passive (non-parallel) antecedent. The passive antecedent is non-parallel because the verbphrase in the active sentence, which is the only plausible antecedent, is no longer a constituent in the passive. Example target sentences with deep and surface anaphors are illustrated in (7c) and (7d), respectively.

Method

Subjects.   A total of 36 undergraduates were recruited from introductory psychology courses.

Materials.   The test materials were constructed from sentence pairs consisting of a context sentence and a target sentence. The sentence pairs were drawn from 20 sets of sentences similar to those in (7). The full set of materials is presented in the Appendix. Each set contained two context sentences, an active (parallel) sentence and a passive (non-parallel) sentence, and two target sentences, a target sentence that ended with a surface anaphor and a target sentence that ended with a deep anaphor. The deep and surface anaphors were chosen so that their interpretations would be similar, if not identical. Within a set, each context set was paired with each target sentence, resulting in four sentence-pairs which were counterbalanced across four presentation lists. The test sentences were intermixed with 39 filler sentences. Sixteen of the filler sentence pairs were constructed so that the target sentence would *not* make sense given the context sentence, as exemplified in (8).

8a. After the exam Bill decided to have a beer or two. Sam didn't either.

8b. Yesterday, the sports star announced his retirement. Sam denied it, too.

Procedure. The sentences were presented using an Apple 2E micro-computer that was interfaced to a Digitry CTS card. Each trial began with a number. The subject then pressed a response button to see the context sentence. When the subject finished reading the context sentence, he or she pressed a response button which erased the context sentence and replaced it with the target sentence. The subject then pressed the appropriate response button to indicate whether or not the target sentence made sense given the context sentence. Examples of sentence-pairs that were drawn from the nonsense fillers were included in the instructions.

Results and Discussion

The subjects' responses yielded two types of data: a "yes" or "no" judgement, and the latency to make the judgement. Separate analyses were conducted on the proportion of sentences judged to make sense, hereafter the *judgement* data, and on the decision latencies for those target sentences that were judged to make sense, hereafter the *latency* data. Note, however, that both the latency and judgement data come from the same set of responses. The latencies are the time that the subject took to make the decision that the target sentence made sense. All latencies that were more or less than two standard deviations above or below a subject's mean were replaced by the appropriate two standard deviation cut-off score, which was calculated for each subject.

The means for the judgement data and the latency data are presented in Table 1. Surface anaphors with non-parallel antecedents were judged to make sense less often than surface anaphors with parallel antecedents, whereas parallelism had little effect on judgements for the deep anaphors. The latency data showed effects of parallelism for both types of anaphors, with the effects being numerically larger for the surface anaphors.

Judgement Analyses. An analysis of variance on the proportion of sentences judged to make sense revealed a main effect of the type of anaphor [$F1(1,28) = 18.39$, $P < 0.01$; $F2(1,19) = 21.00$, $P < 0.01$] and an effect of the parallelism of the antecedent [$F1(1,28) = 16.83$, $P < 0.01$; $F2(1,19) = 9.68$, $P < 0.01$]. The result of primary interest is the interaction between type of anaphor and parallelism [$F1(1,28) = 8.66$, $P < 0.01$; $F2(1,19) = 6.28$, $P < 0.05$]. Separate analyses for the deep and surface anaphors revealed a significant effect of parallelism for the surface

TABLE 1

Percentage of Deep and Surface Anaphors Judged to Make Sense and Judgement Latencies as a Function of the Parallelism of the Context for Experiment 1

	Type of Antecedent			
	Parallel		Non-parallel	
Type of Anaphor	% Judged Sensible	Latency	% Judged Sensible	Latency
Deep	94	2073 msec	91	2273 msec
Surface	89	2161 msec	70	2776 msec

Note: Judgement latencies are reported only for those sentences judged to make sense.

anaphors [$F1(1,28) = 6.14$, $P < 0.01$; $F2(1,19) = 9.42$, $P < 0.01$], but not for the deep anaphors [$F1(1,28) = 0.79$; $F2(1,19) = 1.34$].

Latency Analyses. An ANOVA on the latencies for those sentences judged to make sense revealed an effect of type of anaphor [$F1(1,28) = 7.69$, $P < 0.01$; $F2(1,19) = 6.56$, $P < 0.05$] and an effect of parallelism [$F1(1,28) = 28.33$, $P < 0.01$; $F2(1,19) = 9.38$, $P < 0.01$]. The interaction between type of anaphor and parallelism of the antecedent was not significant [$F1(1,28) = 1.65$; $F2(1,19) = 3.28$].

The results indicate that the comprehension of deep and surface anaphors is affected by the linguistic form in which their antecedents are presented, as suggested by the Sag–Hankamer hypothesis. Surface anaphors were judged to make sense more often when their antecedent was syntactically parallel, whereas syntactic parallelism did not significantly affect judgements to the deep anaphors. However, parallelism did affect comprehension latencies to both types of anaphors. Because similar patterns of results obtained in the latency data for all three experiments, we will postpone considering the implications of this for the Sag–Hankamer hypothesis until the general discussion.

EXPERIMENT 2

In this experiment, non-parallel antecedents were created by presenting the antecedents in a nominalised form. As a result, the non-parallel antecedent is a constituent (unlike Experiment 1), but a constituent of the wrong syntactic category for a verbphrase surface anaphor—a nounphrase rather than a verbphrase. However, the nounphrase should introduce an event into the reader's discourse model, thus creating an appropriate conceptual antecedent for a deep anaphor. An example material set is

presented in (9):

9a. It always annoys Sally when anyone mentions her sister's name.
9b. The mention of her sister's name always annoys Sally.
9c. However, Tom did it anyway out of spite.
9d. However, Tom did anyway out of spite.

The parallel and non-parallel antecedents are presented in the sample context sentences (9a) and (9b), respectively, and the deep and surface anaphors in the sample target sentences (9c) and (9d).

Method

Subjects.   Twenty-eight undergraduates recruited from introductory psychology courses participated in the experiment.

Materials and Procedure.   We used the same lists as in Experiment 1 with the test sentences replaced by sentence-pairs generated from 20 sets of materials similar to the set presented in (9). All of the test materials are presented in the Appendix. The procedure was identical to that used in Experiment 1.

Results

The portion of sentences judged to make sense and the mean latencies for the sentences judged to make sense in each condition are presented in Table 2.

Judgement Analyses.   The judgement results are very similar to those obtained in the first experiment. Surface anaphors were judged to make sense less often when their antecedents were syntactically non-parallel than

TABLE 2

Percentage of Deep and Surface Anaphors Judged to Make Sense and Judgement Latencies as a Function of the Parallelism of the Context for Experiment 2

| | Type of Antecedent | | | |
| | Parallel | | Non-parallel | |
Type of Anaphor	% Judged Sensible	Latency	% Judged Sensible	Latency
Deep	86	2686 msec	86	2954 msec
Surface	89	2556 msec	71	2923 msec

Note: Judgement latencies are reported only for those sentences judged to make sense.

when they were syntactically parallel. However, syntactic parallelism had negligible effects on judgements to deep anaphors. The effect of type of anaphor was nearly significant in the subjects analysis [$F1(1,24) = 3.38$, $P < 0.08$] and significant in the items analysis [$F2(1,19) = 6.72$, $P < 0.05$]. There was a significant effect of parallelism [$F1(1,24) = 8.28$, $P < 0.01$; $F2(1,19) = 8.39$, $P < 0.01$]. The interaction between type of anaphor and parallelism was also significant [$F1(1,24) = 15.61$, $P < 0.01$; $F2(1,19) = 8.39$, $P < 0.01$]. Separate analyses of the surface and deep anaphors showed that the effect of parallelism was significant for the surface anaphors [$F1(1,28) = 16.94$, $P < 0.01$; $F2(1,19) = 14.22$, $P < 0.01$], but not for the deep anaphors (both $F1$ and $F2 < 1.0$).

Latency Analyses.   The latency means showed an increase in latency for both surface and deep anaphors with non-parallel antecedents, which was reflected in an effect of parallelism [$F1(1,24) = 4.54$, $P < 0.05$; $F2(1,19) = 5.51$, $P < 0.05$]. No other effects reached significance (all $Fs < 1.0$).

As in Experiment 1, different patterns of results obtained in the judgement data and in the latency data. Syntactic parallelism had clear effects on whether or not surface anaphors were judged to make sense, but no effects on judgements to deep anaphors. However, when anaphors were judged to be comprehensible, parallelism affected the speed with which both surface and deep anaphors were comprehended.

EXPERIMENT 3

Experiment 3 contrasted null complement anaphora and VP-ellipsis. This contrast is of interest because null complement anaphora is an example of a deep anaphor in which the anaphoric element—the implied complement—is not phonologically realised, as is typically the case for surface anaphors. An example material set is presented in (10):

  10a. Someone has to take out the garbage.
  10b. The garbage has to be taken out.
  10c. But Bill refused.   (Null complement anaphora)
  10d. But Bill refused to.   (VP-ellipsis)

The full set of materials are presented in the Appendix. Both nominal and passive non-parallel antecedents were used.

Method

Subjects.   A total of 48 students were recruited from introductory psychology courses.

Materials and Procedure.   The same filler materials were used as in the first two experiments. Only 12 sets of materials were used because of the limited number of verbs that can be used with null complement anaphora. One of the material sets was later discarded because of a mistake in counterbalancing. Other aspects of the procedure were similar to those described for Experiment 1.

Results and Discussion

The proportion of sentences judged to make sense and the latency to respond to those sentences judged to make sense are presented in Table 3.

Judgement Analyses.   The parallelism of the antecedent had clear effects on judgements to sentences with surface anaphors, but not on judgements to sentences with deep anaphors. There was a significant effect of parallelism both by subjects $[F1(1,40) = 14.44, P < 0.01]$ and by items $[F2(1,10) = 9.52, P < 0.01]$. The effect of type of anaphor was not significant in the subjects analysis $[F1(1,40) = 1.56, P < 0.20]$, but it approached significance in the items analysis $[F2(1,10) = 3.65, P < 0.10]$. The interaction between parallelism and type of anaphor was significant in both the subjects analysis $[F1(1,40) = 5.16, P < 0.01]$ and the items analysis $[F2(1,10) = 5.61, P < 0.05]$. Separate analyses on the surface and deep anaphors showed significant effects of parallelism for the surface anaphors $[F1(1,40) = 14.70, P < 0.01; F2(1,10) = 16.92, P < 0.01]$, but not for the deep anaphors $[F1(1,40) = 1.37; F2(1,10) = 1.00]$.

Latency Analyses.   Parallelism was the only effect to reach significance in the latency analysis [by subjects: $F1(1,40) = 4.41, P < 0.05$; by items: $F2(1,10) = 11.91, P < 0.01$].

TABLE 3

Percentage of Deep and Surface Anaphors Judged to Make Sense and Judgement Latencies as a Function of the Parallelism of the Context for Experiment 3

| | Type of Antecedent | | | |
| | Parallel | | Non-parallel | |
Type of Anaphor	% Judged Sensible	Latency	% Judged Sensible	Latency
Deep	92	1829 msec	89	2147 msec
Surface	95	2023 msec	77	2139 msec

Note: Judgement latencies are reported only for those sentences judged to make sense.

The results of Experiment 3 are similar to those obtained in the first two experiments. The surface anaphors were judged to make sense significantly more often when their antecedent was syntactically parallel, but parallelism did not significantly affect judgements to deep anaphors. Parallelism did, however, influence the latency with which anaphors were judged to make sense, with no interaction between parallelism and type of anaphor, although the parallelism effects were numerically larger for the deep anaphors.

GENERAL DISCUSSION

The experiments reported here were conducted to determine whether the syntactic parallelism of an antecedent has different effects on the processing of deep and surface anaphors. According to the hypothesis developed by Sag and Hankamer (1984), surface anaphors but not deep anaphors must be linked to the linguistic form of their antecedents. The experiments presented here found clear evidence for an interaction between syntactic parallelism and type of anaphor. In all three experiments, surface anaphors that were preceded by syntactically parallel antecedents were judged to make sense more often than surface anaphors preceded by non-parallel antecedents. Syntactic parallelism did not, however, affect judgements to the deep anaphors. This result obtained when syntactic parallelism was manipulated using active–passive contrasts to vary whether the antecedent was a syntactic constituent or not (Experiment 1) and when syntactic parallelism was manipulated by introducing the antecedent in either a verbphrase or a nounphrase constituent (Experiment 2). The same pattern also obtained when both the deep and surface anaphors were not realised phonologically (Experiment 3). Taken together, these results provide clear support for the hypothesis that there is a representational difference between deep and surface anaphors that is reflected in processing.

However, there are two aspects of the data that raise potential problems for the hypothesis that the interpretation of a surface anaphor is necessarily *mediated* by the linguistic representation of its antecedent. The first problem is that surface anaphors remained relatively comprehensible, even when the antecedent was syntactically inappropriate. The second is that syntactic parallelism affected latency judgements to deep anaphors as well as surface anaphors. We will consider each of these aspects of the data in turn.

Although the lack of a syntactically appropriate antecedent interfered with the comprehension of surface anaphors, they were still judged to be comprehensible more than 70% of the time. There are at least two possible explanations for this. The first is that the interpretation of a surface anaphor involves two processes that take place in parallel. One process is

indexing the anaphor with the appropriate syntactic representation and the second is constructing an interpretation making use of discourse model representations. On this view, the linguistic representation of the antecedent of a surface anaphor need not mediate the interpretation of the anaphor. The comprehensibility of a sentence containing a surface anaphor without a syntactically inappropriate antecedent becomes analogous to the comprehensibility of ungrammatical sentences with syntactic violations, such as those illustrated in (11):

11a. *Bill and John was good friends.
11b. *Which friend did John donate the money?

In (11a) the verb does not agree in number with the subject and in (11b) the preposition "to" that would mark the "gap" that "Which friend" is associated with is missing. None the less, both sentences can be interpreted fairly easily despite the fact that both clearly contain syntactic violations. The point is that reconstruction of a full syntactic structure is not necessary for interpretation. Rather, the syntactic form of a sentence is just one aspect of its representation.

An alternative possibility is that the syntactic antecedent of a surface anaphor does, in fact, mediate its interpretation, but that readers were often able to transform the non-parallel antecedent to reconstruct the appropriate syntactic antecedent. On this view, interpreting surface anaphors with non-parallel antecedents should always take longer than interpreting deep anaphors in comparable environments, with the difference being determined by how difficult it is to reconstruct the appropriate antecedent. Although the interaction between parallelism and type of anaphor did not approach significance, we did find numerically larger parallelism effects for surface anaphors than deep anaphors in the first two experiments where the data were based on more than twice as many observations per subject as in Experiment 3. Thus we cannot definitively rule out a reconstruction model. One way of testing such a model would be to directly manipulate how difficult it is for subjects to transform different non-parallel structures into their parallel forms and then determine whether this correlates with parallelism effects for surface anaphors.

The result that is least expected on Sag and Hankamer's hypothesis is that syntactic parallelism consistently affected comprehension times to sentences containing deep anaphors. This result obtained in all three of our experiments (and see Murphy, 1985a).

Murphy accounts for these effects by proposing that both deep and surface anaphors make use of syntactic copying when the syntactic form of the antecedent is available. This explanation runs into several difficulties. It cannot account for the linguistic observations that originally motivated the deep and surface distinction. It also cannot account for the interaction

between parallelism and type of anaphor that obtained in all three of the current experiments.

One possible explanation of these facts proposed by Murphy (1985b) is that while readers and listeners make use of syntactic information in processing both surface and deep anaphors, syntactic information is more important for comprehending surface anaphors because they generally contain fewer "cues" to the nature of their antecedent than do deep anaphors. If this is correct, surface anaphors might be judged to make sense less often than deep anaphors when their antecedents are non-parallel because readers have more difficulty finding the antecedent for the surface anaphor in the discourse model. When an antecedent could not be found within a certain deadline, subjects might have responded "No". A clear test of Murphy's cue hypothesis is provided by Experiment 3, in which we contrasted null complement anaphora with verbphrase ellipsis. In neither of these types of anaphora is the anaphor phonologically realised and on Murphy's analysis both contain the same number of cues. Contrary to the predictions made by the cue hypothesis, syntactic parallelism continued to affect judgements to surface anaphors more than judgements to deep anaphors. Moreover, the size of the interaction between parallelism and type of anaphor in the judgement data was similar in all three experiments.

We are left, then, with the puzzle of why deep anaphors were affected by differences in syntactic structure. One possibility is that contrary to our initial working assumptions, the syntactic structures that we used to introduce parallel and non-parallel antecedents map on to different discourse structures. Clearly, they tend to map on to subtly contrasting logical forms, connected with the truth-conditional semantics of the sentences themselves. Consider the contrast between (12a) and (12b):

12a. Someone needed to paint the garage.
12b. The garage needed to be painted by someone.

Such pairs are widely known to differ in their scoping possibilities. On the favoured reading of (12a), the full propositional content of "someone paints the garage" occurs within the scope of "need". Informally, we might say that the propositional structure of the whole sentence is [Need (someone paint the garage)]. Sentence (12b), however, favours a different scoping of elements. A common observation about passive sentences is that they tend to "focus" on the subject nounphrase, as opposed to their active counterparts, which appear to have their default focus on the main verb, if in fact they have a focus at all. In the semantics of a sentence or discourse, focus tends to be correlated with the wide scoping of the focused element (see Rooth, 1985, for a precise formal hypothesis about the semantics of focus). Thus in (12b), the logical form would be approxi-

mately [The garage$_x$ (Need (someone paint x))]. This corresponds to a situation in which the *state of the garage* itself motivates the need for the painting (e.g. the paint is old and peeling), whereas the semantics for (12a) is more neutral, specifying only that a certain situation needs to come about, though no element is scoped out to target the source of the need.[2] Thus the passive sentence tends to structure the information in such a way as to give special prominence to the subject nounphrase, whereas the active sentence tends not to. If these focus differences are reflected in the discourse model—as they appear to be intuitively—sentence (12b) is more likely to bring to mind an image of a garage with peeling paint than (12a), and if anaphoric processing searches for the most prominent elements of the discourse representation, then the antecedent for a deep verbphrase anaphor which is an event in the discourse model would be less accessible for the passive than for the active. The same may well hold for the comparisons between nominalisations and finite sentences. Nominals tend to take on a presupposed or "backgrounded" character, rather than forming a part of the assertion itself (Kiparsky & Kiparsky, 1971). If so, this would make the content of a nominalisation less accessible for pronominal reference, and could well produce the patterns of results that obtained in Experiment 2.

What is clearly missing in the story that we have been proposing is a precise account of how (semantic) focus and related notions are realised in discourse and how these differing structures interact with the determination of pronominal reference. Without a clearer understanding of these notions, it is hard to adequately control for these facts in manipulating syntactic parallelism; it is quite possible that every syntactic manipulation has a corresponding or at least a correlated discourse effect (Altmann & Steedman, 1988; Crain & Steedman, 1985), so that no effective controls are available. Thus, as is often the case in experiments that manipulate syntactic variables, it is difficult to know whether the variance due to the syntactic manipulation is best explained in syntactic terms. One empirical prediction that can be made even in the absence of a more precise theory, though, is that a context which focuses on the event in a passive sentence should reduce or eliminate parallelism effects for deep but not surface anaphors. Research investigating this possibility is in progress.

In conclusion, our results clearly demonstrate that surface anaphors are sensitive to the form of the their antecedents in a way that deep anaphors

---

[2]Note that one can with some unnaturalness read this sentence as attributing a garage-painting inner drive to someone, though the lack of specificity of the subject and the pragmatic implausibility of someone having garage-painting needs (as opposed to, say, water-drinking needs) militates against this interpretation.

are not. This conclusion differs from that reached by Murphy (1985a; 1985b). However, like Murphy, we did find that parallelism affects the speed with which deep anaphors are interpreted. We have suggested that this is because non-parallel antecedents are typically associated with discourse structures which make the event that serves as the conceptual antecedent for the anaphor less accessible than the discourse structures associated with the parallel antecedents. This suggestion is clearly speculative, though, and it needs to be evaluated empirically. Our data also leave open questions about the degree to which the linguistic antecedents of deep anaphors mediate access to their discourse model antecedents. While unresolved questions remain about how best to capture the differences between deep and surface anaphors, the experiments reported here present clear evidence that the basic distinction argued for by Hankamer and Sag (1976) and Sag and Hankamer (1984) is reflected in comprehension processes.

REFERENCES

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

Anderson, A., Garrod, S. C., & Sanford, A. J. (1983). The accessibility of prenominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*, *35A*, 427–440.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds), *Natural language parsing: Psychological, computational and theoretical perspectives.* Cambridge: Cambridge University Press.

Garnham, A. (1981). Mental models as representations of text. *Memory and Cognition*, *9*, 560–565.

Garnham, A. (1987). Understanding anaphora. In A. W. Ellis (Ed.), *Progress in the psychology of language*, Vol. 3. London: Lawrence Erlbaum Associates Ltd.

Garnham, A. & Oakhill, J. (1987). Interpreting elliptical verb phrases. *Quarterly Journal of Experimental Psychology*, *39A*, 611–627.

Hankamer, J. & Sag, I. A. (1976). Deep and surface anaphor. *Linguistic Inquiry*, *7*, 391–428.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, Mass.: Harvard University Press.

Kiparsky, P. & Kiparsky, C. (1971). Fact. In D. Steinberg & L. Jakobovits (Eds), *Semantics.* Cambridge: Cambridge University Press.

Lesgold, A., Roth, S., & Curtis, M. E. (1979). Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior*, *18*, 291–308.

Malt, B. C. (1985). The role of discourse structure in understanding anaphora. *Journal of Memory and Language*, *24*, 271–289.

Murphy, G. L. (1985a). Processes of understanding anaphora. *Journal of Memory and Language*, 24, 290–303.

Murphy, G. L. (1985b). Psychological explanations of deep and surface anaphora. *Journal of Pragmatics*, 9, 171–198.

Rooth, M. E. (1985). *Association with focus*. University of Massachusetts Ph.D. dissertation distributed by GLSA, University of Massachusetts.

Ross, J. R. (1967). *Constraints on variables in syntax*. Unpublished dissertation, Massachusetts Institute of Technology.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, 2, 437–442.

Sag, I. A. & Hankamer, J. (1984). Toward a theory of anaphoric processing. *Linguistics and Philosophy*, 7, 325–345.

Tanenhaus, M. K. & Carlson, G. N. (1984). Processing deep and surface anaphors. In *Proceedings of the Fourteenth Annual Meetings of the Northeastern Linguistic Society*. Amherst, Mass.: GLSA.

Tanenhaus, M. K., Carlson, G. N., & Seidenberg, M. (1985). Do listeners compute linguistic representations? In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds), *Natural language parsing: Theoretical, psychological, and computational perspectives*. Cambridge: Cambridge University Press.

Wanner, E. (1974). *On remembering, forgetting, and understanding sentences*. The Hague: Mouton.

Webber, B. L. (1981). Discourse model synthesis: Preliminaries to reference. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

APPENDIX: TEST MATERIALS FOR EXPERIMENTS 1–3

The test materials are presented in four sentence-sets. Sentence (a) in each set is the context sentence that introduces a parallel antecedent; sentence (b) is the context sentence that introduces a non-parallel sentence; sentence (c) is the target sentence that contains a deep anaphor; and sentence (d) is the target sentence that contains a surface anaphor.

Materials used in Experiment 1

1a. Someone took the wood out to the shed last night.
 b. The wood was taken out to the shed last night.
 c. Tom told us that Sally did it.
 d. Tom told us that Sally did.

2a. Mike should really change the oil.
 b. The oil should really be changed.
 c. However he/Mike won't take the time to do it.
 d. However he/Mike won't take the time to.
  ("He" was used in context 2a and "Mike" in context 2b.)

3a. John broke the antique vase which belonged to Mrs Jones.
 b. The valuable antique vase which belonged to Mrs Jones was broken by John.
 c. She was furious that he did it.
 d. She was furious that he did.

4a. Someone broke our basement window last night.
  b. Our basement window was broken last night.
  c. Sam thinks he knows who did it.
  d. Sam thinks he knows who.

5a. An architect designed the elaborate conference room.
  b. The elaborate conference room was designed by an architect.
  c. He was paid a lot of money to do it.
  d. He was paid a lot of money to.

6a. Jim and his friends attended the picnic.
  b. The picnic was attended by Jim and his friends.
  c. They were glad that they did it.
  d. They were glad that they did.

7a. We had to replace our broken garbage disposal.
  b. Our broken garbage disposal had to be replaced.
  c. Jim tried to do it, but he failed.
  d. Jim tried to, but he failed.

8a. Al asked Sally to the concert.
  b. Sally was asked to the concert by Al.
  c. For weeks he had been planning to do it.
  d. For weeks he had been planning to.

9a. When Jane returned from work she found that someone had stolen her stereo.
  b. When Jane returned from work, she found her stereo had been stolen.
  c. She had no idea who did it.
  d. She had no idea who.

10a. Somebody shoveled our driveway last night.
  b. Our driveway was shoveled by someone last night.
  c. A neighbor told us that Tom did it.
  d. A neighbor told us that Tom did.

11a. We asked the plumber if he could fix the sink.
  b. We asked the plumber if the sink could be fixed.
  c. He said he would try to do it.
  d. He said he would try to.

12a. Because Mary was busy Tom sent the package.
  b. Because Mary was busy the package was sent by Tom.
  c. He had promised her that he would do it.
  d. He had promised her that he would.

13a. Mary washed the clothes every weekend.
  b. The clothes were washed by Mary every weekend.
  c. Her older sister Sally refused to do it.
  d. Her older sister Sally refused to.

14a. Nobody had ever climbed the mountain before.
  b. The mountain had never been climbed before.
  c. Yesterday my friend Sally did it.
  d. Yesterday my friend Sally did.

15a. Tom hadn't finished the report yet.
  b. The report hadn't been finished yet.
  c. Tom said that he hadn't been able to do it.
  d. Tom said the he hadn't been able to.

16a. Bill was supposed to feed the cat last night.
  b. The cat was supposed to have been fed last night.
  c. He fell asleep and forgot to do it.
  d. He fell asleep and forgot to.

17a. My brother Sam wrote a best-selling novel.
  b. A best-selling novel was written by my brother Sam.
  c. Our family was proud he did it.
  d. Our family was proud he did.

18a. Somebody had to paint the garage.
  b. The garage had to be painted.
  c. Finally my younger sister Carol agreed to do it.
  d. Finally my younger sister Carol agreed to.

19a. The police are curious about who robbed the jewelry store.
  b. The jewelry store robbery is a puzzle to the police.
  c. They still don't know who did it.
  d. They still don't know who did.

20a. At the sports event the opera star sang the national anthem.
  b. At the sports event the national anthem was sung by an opera star.
  c. People were surprised he agreed to do it.
  d. People were surprised he agreed to.

Materials used in Experiment 2

1a. It would be good for the country if someone would overthrow the oppressive regime.
  b. The overthrow of the oppressive regime would be good for the country.
  c. But nobody has the courage to try to do it.
  d. But nobody has the courage to try to.

2a. It would do you good to jog into town.
  b. A jog into town would do you good.
  c. Please let me know if you decide to do it.
  d. Please let me know if you decide to.

3a. It took the chef only half an hour to prepare the meal.
  b. The preparation of the meal took the chef only half an hour.
  c. We thought it would take him much longer to do it.
  d. We thought it would take him much longer to.

4a. The CIA agent exposed some important secrets to an enemy agent.
  b. The CIA agent's exposure of some important secrets to an enemy agent was unfortunate.
  c. The CIA is extremely upset he did it.
  d. The CIA is extremely upset he did.

5a. The judge was surprised when the mobster pleaded guilty.
  b. The mobster's plea of guilty was surprising to the judge.
  c. No one had expected he would do it.
  d. No one had expected he would.

6a. It would be dangerous to climb the steep mountain without an experienced guide.
  b. A climb up the steep mountain would be dangerous without an experienced guide.
  c. Nevertheless, my best friend Bill agreed to do it.
  d. Nevertheless, my best friend Bill agreed to.

7a. It could take three hours to swim to the island.
  b. A swim to the island could take three hours.
  c. That is why we decided not to do it.
  d. That is why we decided not to.

8a. Tom knew that threatening resignation would have little effect on the President's policy.
  b. Tom knew that his threat of resignation would have little effect on the President's policy.
  c. But he finally decided he would do it anyway.
  d. But he finally decided he would anyway.

9a. Mary criticised her boss constantly.
  b. Mary's constant criticism of her boss was annoying.
  c. We didn't know why she felt the need to do it.
  d. We didn't know why she felt the need to.

10a. Many accidents this year resulted from bartenders selling alcohol to minors.
  b. The sale of alcohol to minors has resulted in many accidents this year.
  c. Despite stiff fines, many bartenders continue to do it.
  d. Despite stiff fines many bartenders continue to.

11a. It always annoys Sally when anyone mentions her sister's name.
  b. The mention of her sister's name always annoys Sally.
  c. However, Tom did it anyway, out of spite.
  d. However, Tom did anyway, out of spite.

12a. Some unnamed terrorists robbed the Bank of America as an act of desperation.
  b. The robbery of the Bank of America by some unnamed terrorists was a surprise.
  c. The police still haven't figured out who did it.
  d. The police still haven't figured out who did.

13a. The office manager banned smoking in the lounge.
  b. Morale is at a low because of the office manager's ban on smoking.
  c. He is beginning to regret that he did it.
  d. He is beginning to regret that he did.

14a. The country was shocked when someone assassinated the young president.
  b. The country was shocked by the assassination of the young president.
  c. People couldn't understand why anyone would do it.
  d. People couldn't understand why anyone would.

15a. It would be helpful to Bill if someone discussed his poor study habits with him.
  b. Discussion of Bill's poor study habits would be helpful to him.
  c. But I certainly don't have the nerve to do it.
  d. But I certainly don't have the nerve to.

16a. The financial world was stunned when Mr Smith purchased Exxon Oil.
  b. Mr Smith's purchase of Exxon Oil was stunning news to the financial world.
  c. Nobody thought he was rich enough to do it.
  d. Nobody thought he was rich enough to.

17a. The hostage crisis might end if someone would raid the building.
  b. The hostage crisis might end with a raid on the building.
  c. But we need someone with the courage to do it.
  d. But we need someone with the courage to.

18a. It can often be embarrassing to rely on others for help.
  b. Reliance on others for help can often be embarrassing.
  c. However, there are times when we have to do it.
  d. However, there are times when we have to.

19a. In this state it is illegal for politicians to accept gifts.
 b. In this state the acceptance of gifts by politicians is illegal.
 c. Many small town politicians still continue to do it.
 d. Many small town politicians still continue to.

20a. The troops were in dismay when the general surrendered.
 b. The general's surrender was dismaying to the troops.
 c. He had promised them that he never would do it.
 d. He had promised them that he never would.

Materials used in Experiment 3

1a. Someone has to take the wood out to the shed tonight.
 b. The wood has to be taken to the shed tonight.
 c. Last night, Sally forgot.
 d. Last night, Sally forgot to.

2a. John almost didn't turn off the oven last night.
 b. The oven almost wasn't turned off last night.
 c. We were lucky Sally reminded him (John).
 d. We were lucky Sally reminded him (John) to.
    ("Him" was used in context 2a and "John" in context 2b.)

3a. Someone was supposed to feed the cat last night.
 b. The cat was supposed to have been fed last night.
 c. Bill didn't remember.
 d. Bill didn't remember to.

4a. Nobody picked Bill's brother up at the airport.
 b. Bill's brother wasn't picked up at the airport.
 c. Bill couldn't persuade any of his friends.
 d. Bill couldn't persuade any of his friends to.

5a. On Saturday night we needed someone to watch our baby daughter.
 b. Our baby daughter needed to be watched on Saturday night.
 c. My sister offered.
 d. My sister offered to.

6a. John attempted to quit smoking for the third time.
 b. John attempted quitting smoking for the third time.
 c. But, he failed again.
 d. But, he failed to again.

7a. Sally's mother wanted to chaperone the party.
 b. A chaperone was needed for the party.
 c. But she (Sally's mother) wasn't asked.
 d. But she (Sally's mother) wasn't asked to.
    ("She" was used in context 7a and "Sally's mother" in context 7b.)

8a. Jim has to replace his broken garbage disposal.
 b. The broken garbage disposal needs to be replaced.
 c. He (Jim) will try this weekend.
 d. He (Jim) will try to this weekend.
    ("He" was used in context 8a and "Jim" in context 8b.)

9a. Someone had to paint the garage.
 b. The garage had to be painted.
 c. My friend Carol voluntecred.
 d. My friend Carol volunteered to.

10a. Tom said he hadn't finished the report yet.
  b. Tom said the report hadn't been finished yet.
  c. He hadn't had a chance.
  d. He hadn't had a chance to.

11a. We asked the plumber if he could fix the sink.
  b. We asked the plumber if the sink could be fixed.
  c. He said he would try.
  d. He said he would try to.