Processing of Deep and Surface Anaphors

A Replication and Extension of Tanenhaus, Carlson & Seidenberg (1985)

Suzanne Bélanger
University of Toronto

Tanenhaus, Carlson, & Seidenberg (1985) found striking evidence that deep and surface anaphors, as described in Hankamer & Sag (1976) and Sag & Hankamer (1984), access different levels of representation during resolution. We argue that a full interpretation of their results is not possible without control conditions. Specifically, it is not clear whether surface anaphors affect the linguistic representation by keeping it active or whether deep anaphors affect the linguistic representation through suppression. We ran two experiments intended to replicate their main finding and also to shed light on the activation/suppression issue. We failed to replicate their result in Experiment 1. Our second experiment controlled for possible confounds, and replicated their result. Moreover, the results supported the hypothesis that suppression is involved with deep anaphors. Further research is required, however, to determine whether surface anaphors affect the level of activation of the linguistic representation or not.

1. Introduction

In the anaphora literature, there has been considerable discussion of whether or not different types of anaphors access different levels of representation during resolution (Sag & Hankamer, 1984; Cloitre & Bever, 1988; Murphy 1985a, b; Carlson & Tanenhaus, 1984; Tanenhaus, Carlson & Seidenberg, 1985; Murphy, 1990; Tanenhaus & Carlson, 1990; Mauner, Tanenhaus & Carlson, 1995). Given the various forms of anaphora available for establishing coherence in discourse (pronouns, ellipses, definite NPs, to name a few), it is reasonable to ask whether anaphora resolution is essentially uniform, regardless of anaphor type, or whether the process differs depending on the type of anaphor involved.

Hankamer & Sag (1976) and Sag & Hankamer (1984) proposed that anaphors can be divided into two classes: ‘deep’ and ‘surface’. Included in the class of deep anaphors are
expressions such as ‘do it/do the same thing/do that’, as well as personal pronouns (he, she, it, they, etc.), ‘sentential it’, one-anaphora and null complement anaphors. Surface anaphors include verb phrase ellipsis, sluicing, gapping, stripping and so-anaphors.

(1) Deep:

Merle smashed the red Cadillac with a sledgehammer.

a. She did it because she wanted to.  (do-it)
b. She did it because she wanted to.  (pronoun)
c. It was good exercise.  (sentential it)
d. She did it whenever we let her ∅.  (null complement anaphor)
e. Betty smashed the green one.  (one-anaphor)

(2) Surface:

a. John jumped off the cliff and Pete did too.  (VPE)
b. John jumped off the cliff but we don’t know why.  (sluicing)
c. John likes berries and Pete, pears.  (gapping)
d. I’ll take you to the movies but not this week.  (stripping)
e. You say she’s guilty but do you really think so?  (so-anaphor)

The justification for the two classes comes from the fact that, as H&S point out, there exist certain formal/distributional differences between them. Deep anaphors, for instance, do not require a linguistic antecedent, whereas surface anaphors generally do. This is illustrated in the following example taken from S&H, 1984:

(3) [Hankamer points gun offstage and fires, whereupon a bloodcurdling female scream is heard]

Sag:

a. *I wonder who?  (sluicing)
b. *I wonder who was?  (VPE)
c. *Jorge, you shouldn’t have.  (VPE)
d. I wonder who she was?  (pronoun)
e. Jorge, you shouldn’t have done it.  (do-it)

1 Hankamer & Sag (1976) and Sag & Hankamer (1984) actually collapse these types of anaphors with ‘S-anaphors’ or ‘sentential it’. Subsequently, Murphy (1985a) makes a distinction between sentential it and do it/that, etc., the latter being considered more verbal than sentential. Tanenhaus, Carlson & Seidenberg (1985), Tanenhaus & Carlson (1990) and Mauner, Tanenhaus & Carlson (1995) also classify ‘do it’ as a verbal anaphor. The following illustrates this distinction:

   (i) Sean made a face at the conductor.
      a. It really bothered his mother. (it= Sean making a face at the conductor)
      b. His mother was mad that he did it. (it=made a face at the conductor).

2 Later more types of anaphors were classified as being deep or surface. For instance, Hestvik, Nordby, Karlsen & Pedersen (1998) add reflexives and traces to the class of surface anaphors. It is certainly true that both of these require a linguistic antecedent—traces are, however, a theory-internal construct whereas reflexives are not.
Surface anaphors also require that their antecedent be in some sense ‘parallel’, whereas deep anaphors do not. This is shown in sentence (4) where an active surface anaphor, as in (4b), cannot take a passive antecedent:

(4) The children asked to be squirted with the hose so
   a. they were (VPE)
   b. *we did (VPE)
   c. we did it. (do-it)

H&S argue that formally, deep and surface anaphors are fundamentally different. Deep anaphors are present at deep structure, whereas surface anaphors are not. Rather, surface anaphors arise via a deletion operation. At the level of deep structure, then, an elided construction such as (5a) would be represented as in (5b):

(5) a. John went to school and Bill did too.
   b. John went to school and Bill [went to school] too.

Sag (1977) argues that the deletion operation is subject to an identity condition which applies at the level of logical form. At this level, the two VPs are assessed, and the second VP is deletable under identity with the first. The problem with non-parallel antecedents such as in (4) is that the two VPs fail to meet the identity requirement at LF, and so deletion is not permitted.3

Later Sag & Hankamer (1984) propose that the formal differences between deep and surface anaphors have a processing correlate: deep anaphors access their antecedents in the ‘mental model’ (a non-linguistic level) whereas surface anaphors access their antecedents at the level of logical form, a linguistic level.

Certain predictions can be made from the simple linguistic/non-linguistic distinction proposed by S&H. The fact that deep anaphors can take a pragmatic antecedent whereas surface anaphors cannot might be taken as support for the notion that different levels might be involved in processing. Further, S&H note that surface anaphors appear to be affected by distance from their antecedents, whereas deep anaphors do not. An example of this is as follows.4

(6) John raked the leaves in the back yard.
   This was much more fun than studying for exams.
   ?Later, Bill did too. (surface)
   Later, Bill did it too. (deep)

---

3 An alternative view presented in Williams (1977) is that the ellipsis is represented as an empty category at deep structure and receives its interpretation at logical form through a kind of reconstruction operation which copies in the relevant VP. The differences between these analyses are not important here. Both approaches make reference to logical form, not surface structure, as the relevant level.

4 Note that the reduced acceptability of the ellipsis here does not appear to be a grammaticality issue—the ellipsis is grammatical when the antecedent is not in the immediately preceding sentence, just somewhat less felicitous.
It is well known that surface information (exact words and word order information) decays very rapidly (Sachs, 1967, 1974). If it is true that surface anaphors require access to this rapidly decaying representation, these anaphors will need to be very close to their antecedents, and processing will suffer with increased distance. A related prediction not discussed by S&H is that one might also expect that aspects of surface structure that may detrimentally affect surface anaphors if close (such as lack of parallelism) should have their effects attenuated with distance. This prediction is investigated experimentally in Murphy (1985a), Experiment 3, and is discussed below.

Although it is reasonable to posit that the formal differences between deep and surface anaphors have a processing correlate, it is by no means necessarily the case. Deep anaphors may access the mental model or conceptual level directly, or they may do so via a linguistic antecedent, insofar as it is available. If the latter scenario holds, then resolution with a linguistic antecedent would proceed (roughly) the same for deep and surface anaphors. I say ‘roughly’ because it is unclear a priori whether accessing a linguistic representation would take the same form or have the same processing consequences for both deep and surface anaphors—it is possible that deep and surface anaphors use the linguistic representation in different ways. A related question (Ron Smyth, p.c.), which we will not examine here, is whether we would expect longer resolution times for a deep anaphor that has a linguistic antecedent compared to one that finds its antecedent directly in the mental model.

Since S&H (1984), many researchers have investigated the deep-surface processing distinction experimentally. I will focus here on the ones that have compared ellipsis (a surface anaphor) with the deep anaphor ‘do it’. These studies can be divided into roughly two classes: those that manipulate distance between antecedent and anaphor (Murphy, 1985a; Tanenhaus, Carlson & Seidenberg, 1985; Murphy, 1990) and those that manipulate the form of the antecedent (Murphy 1985a; Tanenhaus, Carlson & Seidenberg, 1985; Tanenhaus & Carlson, 1990; and Mauner, Tanenhaus & Carlson, 1995). These studies are reviewed in the following sections. In addition, there has been considerable research on resolution of other anaphors (pronouns, definite NP anaphors, reflexives, traces, PRO) which contributes to the deep-surface debate. I will discuss the relevance of some of these studies in the final section of this paper.

Following the literature review I report on two experiments. The goal of these experiments was to replicate and extend findings in Tanenhaus, Carlson & Seidenberg (1985). This study found perhaps the most striking and, if true, compelling, evidence for differential processing of deep and surface anaphors. However, they did not provide any statistical tests or details regarding subjects or procedure. Further, we believe that their materials were lacking the proper control conditions required to fully interpret the results.

2. Effects of distance

Several studies have investigated the effect of distance on resolution of deep and surface anaphors. These experiments are based on the robust finding that syntactic information decays rapidly, whereas conceptual/semantic information remains relatively stable (Sachs, 1967; 1974). If surface anaphors require access to linguistic/syntactic
information, they should become increasingly difficult to process with distance from the antecedent. On the other hand, deep anaphors, by hypothesis, require access to a conceptual representation, and so resolution should be unaffected by distance.

2.1 Murphy (1985a), Experiments 1 and 2

Murphy (1985a) reports on three experiments involving three factors: length of antecedent, distance of antecedent from anaphor and parallelism of antecedent (passive vs. active). The inclusion of the first factor is based on the assumption that length of the antecedent is a property that is only evident at the linguistic level, not at the conceptual level. If surface anaphors access a linguistic representation, they will be more affected by length than deep anaphors. It is also based on Murphy’s assumption that length will affect resolution if resolution involves a copying procedure, i.e., copying the full antecedent into the anaphor site.

In Experiment 1, Murphy investigated the effect of length of the antecedent on processing of the anaphor (he does not vary distance or parallelism in this experiment). The following is a sample sentence; length was varied by either including or excluding the portion in square brackets:

(7) Jimmy swept the [tile] floor [behind the chairs free of hair and cigarettes.]
   a. Later his uncle did too.
   b. Later his uncle did it too.

The dependent variable was the reading time for the anaphor sentence. The results showed that both deep and surface anaphors were affected by antecedent length (reading times for surface and deep anaphors were slowed in the long condition by 264 msecs and 225 msecs respectively); no interaction between length and form of anaphor was found. This result suggests either that both deep and surface anaphors access a linguistic level and are thus affected by linguistic length, or that they access different levels and are affected by length in different ways. That is, it could be that surface anaphors are affected by linguistic length whereas deep anaphors are affected by the additional information contained in the longer antecedent (but at a conceptual level). Murphy essentially adopts the former interpretation.

---

5 Although Sag & Hankamer stress that the relevant level is logical form, we will continue to use the terms superficial/surface/linguistic to refer to the level hypothesized to be relevant only for surface anaphors. The crucial thing to remember is that the proposed level is a linguistic one for surface anaphors and a non-linguistic one for deep anaphors.

6 Tanenhaus & Carlson (1990) claim that ‘copying is not the mechanism by which a surface anaphor would be associated with its antecedent in most linguistic theories. The more standard assumption is that anaphors are co-indexed with their antecedents, i.e., linked by a pointer to the antecedent.’ (p. 261). It is unclear what it means, in processing terms, to be linked to something by a ‘pointer’ and why one would predict the anaphor to be affected by the form (as Tanenhaus & Carlson do in their paper) of the antecedent but not the length. In addition, since then, many recent theoretical analyses of ellipsis do involve a copy of the antecedent VP at the ellipsis site at some point in the derivation (Fiengo & May, 1994; Chomsky, 1995, among others).
The second experiment introduced the distance variable. As only the following example is provided in the description of the materials, it is unclear what kind of intervening sentences were used:

(8) Antecedent (long or short)
    Intervening sentence
    Anaphoric sentence (deep or surface).

Murphy found a main effect of distance when compared with the results of Experiment 1: reading times were increased for both deep and surface anaphors (441 and 697 msecs. respectively) However, he does not provide any statistics on the interaction of anaphor type with distance across the two experiments. That is, even though both deep and surface anaphors were affected by distance, we do not know whether surface anaphors were more affected than deep (in fact, numerically, surface were more affected by 256 msecs) as would be predicted by the Deep-Surface hypothesis of Sag & Hankamer (1984).

Furthermore, as no materials are provided, we do not know what effect the nature of the intervening sentence might have had on the results. It is possible, for example, that a topic shift was introduced, which might have affected both types of anaphors (see Malt, 1985 for the effect of topic shift on ellipsis).

One interesting result is the interaction of length with distance. Length had an effect on reading times when the antecedent was close, but not when the antecedent was separated from the anaphor by an intervening clause. Moreover, this was true for both deep and surface anaphors. Murphy suggests that this result is consistent with the idea that a linguistic representation of the antecedent is accessed when the anaphor is close to the antecedent but a more abstract representation is accessed when far.

However, it seems at least possible that the additional information contained in the longer antecedent might also affect the complexity of representation in the mental model. Therefore we do not know whether the effect of length in Experiment 1 was in fact due to both anaphors accessing a surface representation. Still, this would not explain why neither surface nor deep anaphors were affected by length when the antecedent was distant. This finding seems to support Murphy’s suggestion that both anaphors were accessing more abstract representations in this condition because the linguistic representation was less available (for both). On the other hand, to the extent that it is possible that length would affect both a linguistic representation and a conceptual one, it is also possible that the effects of length would be attenuated with distance for both. J. Simner (p.c.) suggests the possibility that ‘deep representations’ might become more compact with distance.

Nevertheless, in the absence of any statistical tests on the interaction of interest, we are left with no reliable results on the differential effect of distance on deep versus surface anaphors.

2.2 Tanenhaus, Carlson & Seidenberg (1985)

Tanenhaus, Carlson & Seidenberg (1985) also tested the effect of distance on deep and surface anaphors. They conducted an experiment similar to Murphy’s, but instead of using VPE as a surface anaphor, they used sluicing. An example of their materials is
shown below; the portion in square brackets was included in the ‘distant’ condition and excluded in the ‘close’ condition:

(9) Somebody has to paint the garage.
    [The paint is peeling and the wood is beginning to rot.]
    Let’s take a vote and see who has to do it. (deep)
    Let’s take a vote and see who. (surface)

Although no statistics are reported, the results indicate that in the distant condition, surface anaphors are read 207 msecs slower than in the close condition, whereas deep anaphors are in fact read more quickly by 84 msecs. This finding appears to support S&H’s hypothesis that deep and surface anaphors access different levels of representation.

However, there is a confound in their materials (also noted by Murphy 1990): the intervening sentence provides an alternative antecedent for the surface anaphor, but not for the deep anaphor. Where as ‘do it’ anaphors cannot take a stative antecedent, sluicing can, as in the following:

(10) Let’s take a vote and see who [is peeling and beginning to rot].

While pragmatically odd, this interpretation is nevertheless perfectly grammatical. This difference between deep and surface anaphors could be responsible for the differences observed. Surface anaphors would take longer if the closer, but implausible, antecedent is canvassed before the more distant one is considered. In fact, as no follow-up questions are reported, we do not know how subjects ultimately interpreted the anaphor. Consequently, we still have no clear result on any differential effect of distance on deep vs. surface anaphors.

2.3 Murphy (1990) – Comparison of Tasks

Murphy (1990) observed the conflicting results of the studies discussed above, and also noted the potential confound in Tanenhaus et al.’s experiment. He conducted two studies designed to further test the effect of distance and to investigate the possible effect of ambiguity. He also used the two experiments to test the effect of task on results.

For both experiments, he included three factors: distance (close vs. distant), ambiguity (distant condition only; ambiguous vs. unambiguous), and anaphor type (VPE vs. ‘do it’). He constructed texts that contained an antecedent and two types of intervening sentence for the distant condition: an ambiguous one that provided an alternative antecedent for the VPE, and an unambiguous one which did not. Note that the ambiguous sentence was only expected to affect VPE, as these sentences did not provide a grammatical antecedent for ‘do it’. The close condition had no intervening sentence. The following example is provided by Murphy:

(11) Ellen Marcovitz was flying from Seattle to Washington, D.C.
    She was a sales representative for Acme Aviation, Inc.,
and she was trying to get a big government contract for her company.
Because she was nervous, she started a conversation with the man next to her.
The man asked her a question about Acme Aviation.  
–This was more relaxing than her previous worrying. 
–She felt more relaxed almost immediately.  
Later, Ellen wondered why he did/did it.
She was worried that she might have told something important to this stranger.

As noted, the ambiguous intervening sentence was intended to provide an alternative, syntactically appropriate antecedent for the ellipsis only. So, ‘Later she wondered why he did’ could be interpreted as ‘Later she wondered why he [felt more relaxed almost immediately].’ Murphy claims that this interpretation is implausible and that the intended antecedent is the ‘correct’ antecedent according to pragmatic reasoning. If ambiguity has an effect, reading times for VPE clauses will be slower in the distant-ambiguous condition than in the distant-unambiguous condition, whereas no difference is expected for ‘do-it’.

In addition, texts were followed by a true-false question in Experiment 1, but not in Experiment 2 (discussed below).

In Experiment 1, the dependent variable was reading time for the anaphor clause. The results of this experiment indicate that neither VPE nor ‘do-it’ were affected by ambiguity—in fact, the reading times for the ambiguous condition were slightly faster than the unambiguous condition for both anaphors.

However, a comparison between Murphy’s sample of his materials and Tanenhaus et al.’s reveals some differences. First, in Tanenhaus et al.’s example, there is no disambiguating factor other than plausibility, whereas in Murphy’s, the subject pronoun in the anaphor sentence (‘he’ in the above example) serves to assist in disambiguating the intended reading, and might even act as a pointer to the intended antecedent verb phrase since the subject of that VP is the antecedent of the pronoun. This might have the effect of collapsing any differences that might arise due to ambiguity.

Second, Murphy himself suggests that subjects might have ‘chosen’ the ambiguous sentence as the antecedent, even though it was implausible. In fact the degree of implausibility appears to differ across the two studies. The implausible antecedent is quite implausible in Tanenhaus et al.’s study (recall that it was not intended to be an alternative antecedent for the surface anaphor). In Murphy’s, it depends on how subjects are reading the anaphor clause. If they read it with stress on ‘he’, then the clause makes sense with the implausible but closer antecedent. That is, it is not completely unreasonable for Ellen to wonder why the man was made more relaxed by his asking a question about her company. The stress reading for the surface anaphor might provide a plausible reading with the immediately preceding VP as antecedent, effectively negating the need to canvas any additional antecedents (but nevertheless increasing reading time relative to the close condition because of the marked reading).

In short, Murphy’s Experiment 1 does not really test whether ambiguity of the kind found in Tanenhaus et al.’s study affected the results of that study. Thus we cannot rule out ambiguity as a factor in the Tanenhaus et al. study. The two experiments also employed different kinds of surface anaphors (Murphy used verb phrase ellipsis whereas Tanenhaus et al. used sluicing), which makes any comparison between them all the more difficult.
As for the effect of distance on deep versus surface anaphors, in Experiment 1, Murphy again found no difference in the extent to which deep and surface anaphors were affected by distance - surface anaphors were slowed by 307 msecs and deep anaphors, by 292 msecs. The interaction of anaphor type with distance was non-significant (both Fs < 1.0).

Noting that other researchers investigating the deep-surface difference have used a ‘makes-sense’ judgment task as opposed to a reading time task and have found a deep-surface processing difference (these studies will be discussed below in the next section), Murphy set out in Experiment 2 to explore the effects of task on results.

For this experiment, he used the same materials as for Experiment 1, with some additional items. Subjects read passages and had to indicate whether or not the anaphor sentence made sense given the preceding context. The anaphor sentence was always a grammatical sentence; nonsensicality arose from referring to or presupposing things not mentioned in the preceding text (unfortunately no examples are provided). Because the task required subjects to make a judgment on the anaphor sentence, the true-false questions were not given to subjects in this study.

The factors were the same as in Experiment 1 (distance, ambiguity and anaphor type). The two dependent variables were judgment latencies (latency data) and the judgment accuracy (judgment data). Judgment accuracy reflects judgments made on anaphor sentences that made sense. Table 1 indicates percentages of ‘No’ judgments to sensible anaphor sentences and the times taken to render judgments to target sentences, which were all sensible (times included both positive and negative judgments).

<table>
<thead>
<tr>
<th></th>
<th>Close</th>
<th>Distant-Unambiguous</th>
<th>Distant-Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>Percent responses</td>
<td>No</td>
<td>Times</td>
</tr>
<tr>
<td>Deep</td>
<td>Percent responses</td>
<td>9</td>
<td>2019</td>
</tr>
</tbody>
</table>

Table 1: Percent “No” responses (judgment data) and Mean Judgment Latencies (latency data) for Murphy (1990) Experiment #2

The results for the judgment data indicate that there is little effect of ambiguity (distant-ambiguous vs. distant unambiguous). There was an effect of distance, with more negative judgments on the anaphor sentences that made sense associated with the distant-unambiguous condition than the close condition. However, there was no effect of anaphor type. Both deep and surface anaphors were negatively affected by distance from the antecedent in terms of the judgment data. The interaction of distance and anaphor type was not reliable (both Fs<1.0).

The latency data, however, do show an interaction of distance and anaphor type. Surface anaphors showed an increase in reaction time from the close to the distant-unambiguous to the distant-ambiguous condition, whereas deep anaphors actually showed a decrease along these comparisons (Murphy does not provide the p-values for each of these comparisons).

Murphy combines the results of the two studies adding Experiment as a new factor. If the patterns found in the two experiments are reliable, then a three-way interaction
should arise of anaphor, distance and experiment. This interaction was in fact found in the
items analysis, but not in the subjects analysis (p < .15). Nevertheless, because the
interaction of anaphor and distance was significant by subjects for Experiment 2’s latency
data, Murphy tentatively concludes that deep and surface anaphors are differentially
affected by the antecedent’s distance in a judgment task but not in a simple reading task.

Murphy suggests that the different results in reaction times between the two
experiments are due to the differences in task. He proposes that with a simple reading
time task, subjects are not required to specifically evaluate a given sentence, only to
integrate it with what has come before. On the other hand, in a makes-sense task, subjects
are attending to the anaphor sentence on a conscious level, and may be paying more
attention to the formal properties that differentiate deep from surface anaphora. Murphy
says that times in the simple reading time task are affected by the accessibility of the
antecedent (close or distant), not by the formal differences between the anaphors,
whereas the times for a task which requires the subject to make an evaluation will be
affected by the latter. This is consistent with recent work showing that the working
memory system used in syntactic, automatic processing is distinct from the system that
underlies the accomplishment of verbally-mediated tasks involving conscious, controlled
processing (see Caplan & Waters, 1999).

However, this explanation does not readily account for the results of Tanenhaus et
al.’s 1985 experiment on distance discussed above. This study, like Murphy (1985a)
Experiments 1 & 2, used a comprehension time task, which, according to Murphy, should
have shown both deep and surface anaphors to be affected by distance. The ambiguity
confound in the Tanenhaus et al. experiment remains unresolved; however, setting this
confound aside, there is another possible reason for the different pattern of results.
Murphy (1990) indicates that the Tanenhaus et al. study used follow-up questions that
focused on the interpretation of the anaphor (although use of such questions is not
reported in Tanenhaus et al. 1985). Murphy on the other hand used follow-up questions
that did not always focus on the anaphoric sentence; his reasoning for this was that he did
not want subjects to adopt unnatural processing strategies in reading the anaphors. It is
possible that use of questions that always referred to the anaphoric sentence may have
had the same effect as the makes-sense task, encouraging more conscious processing of
the anaphor.

Murphy proposes a single model for the interpretation of both deep and surface
anaphora: 1. the VP anaphor initiates a search through the recent linguistic context for
possible antecedents 2. other antecedents may be located in the discourse model 3. all
candidate antecedents are subjected to a plausibility test. He assumes that ‘the two
sources are in competition, but the exact decision rule listeners use in choosing between
them is not fully known’ (p. 689).

Essentially, then, both types of anaphor are resolved using a ‘two-source process’
(p. 691). Because both deep and surface anaphors initiate a search through the recent
linguistic context for an antecedent, both will be affected by the absence of a suitable
antecedent in that context, in particular, in the most recent clause, which is represented
more strongly in memory than earlier clauses (see, e.g., Jarvella, 1971). According to
Murphy, the reason for the slower response times in the distant condition is that
pragmatic reasoning (which, Murphy claims, may take more time than syntactic
processing) is used to locate an antecedent in the discourse model for both anaphors.
Thus he does not locate the slower reading times in the fact that the linguistic antecedent is far away and subjects struggle to access it (as is suggested by Garnham & Oakhill, 1987 for surface anaphors). Rather, the slower reading times are due to pragmatic reasoning being a more time-consuming antecedent-locating process.

The second point noted above states that both deep and surface anaphors may find their antecedent in the discourse model. This seems to imply that degree of linguistic distance from the antecedent will not have an effect; if the immediate linguistic context is canvassed and an antecedent is not found, both deep and surface anaphors may find their antecedent in the discourse model. But consider the following dialogues (the first from S&H, 1984):

(12)

a. E: Good morning. I would like for you to reassemble the compressor...I suggest you begin by attaching the pump to the platform...
   A: All right. I assume the hold in the housing cover opens to the pump pulley rather than to the motor pulley.
   E: Yes, that is correct. The pump pulley also acts as a fan to cool the pump.
   A: Fine. Thank you.
   ...All right, the belt housing cover is on and tightened down.

30 minutes and 60 utterances from the beginning
   E: Fine. I knew you could do it. [reassemble the compressor] (deep)
   E: Fine. ???I knew you could. [reassemble the compressor.] (surface)

b. A: Jerry had an affair with Susan.
   B: Really? I hate hearing about people cheating.
   A: Apparently they’ve been having problems for a while.
   B: I didn’t know about that.
   A: Yes. And he’s always been a flirt.
   B: Well he’s very charming.

   A: Yes. Still though I’m surprised that he did it. (deep)
   A: Yes. ???Still, though, I’m surprised that he did. (surface)

Now compare this to Murphy’s materials:

(13) The man asked her a question about Acme Aviation.
   This was more relaxing than her previous worry.

   ?Later she wondered why he did.
   Later she wondered why he did it.

Distance of the kind illustrated in (12a) and (12b) has not been tested, the rationale being that surface form decays very rapidly (Sachs, 1967; Jarvella, 1971, among many others) and that one intervening clause would have the effect of ensuring that the surface representation of the antecedent is no longer accessible (or at least not easily accessible).
However, there appears to be a marked difference in the interpretability for the surface anaphor when compared to (13), but not for the deep.

Consider (12a). In this discourse, there is a considerable linguistic distance and time delay from the original utterance to the anaphor. The surface anaphor is quite bizarre, but the deep anaphor is perfectly acceptable. The linguistic representation would surely no longer be available at this point, so both anaphors would find no available antecedent in the recent linguistic context. If both are resolved in the same way, then both should find their antecedents in the discourse model; however, it seems that only the deep anaphor is interpretable, contrary to what Murphy’s model predicts.

One might argue that it is the whole 30 minute event and the successful conclusion that acts as the deictic referent for the deep anaphor, not the mental representation of the initial utterance. In (12b), however, the antecedent is not deictic, and yet the deep anaphor is still far more acceptable. This is expected if only deep anaphors can readily find their antecedent in the discourse model, but not if both may do so. It is worth noting that subjects may, in order to interpret surface anaphors in such cases, treat the surface anaphor as if it were a deep one (see Shapiro & Hestvik, 1995), thereby collapsing any differences that the manipulation might be intended to uncover.

A number of factors could cause a slowdown. Searching the recent clause and finding no antecedent there would increase response times. Employing pragmatic reasoning to locate an antecedent in the discourse model might also take more time than locating the antecedent in the linguistic representation. Struggling to access a distant antecedent would also affect response times. So it is unclear whether deep and surface anaphors are affected by distance for the same reason or combination of reasons. It may be that all of these factors affect both but are weighted differently.

Two things remain unclear about Murphy’s Experiment 2 (the makes-sense experiment), given the model and his explanation of the differences in task. First is the fact that surface anaphors were more affected by distance in the judgment latencies than deep anaphors. Second is the fact that negative judgments for deep anaphors increased in the distant condition. Recall that the analysis of the judgment data revealed that both deep and surface anaphors were adversely affected by distance, but that only surface anaphors showed increased latencies. Murphy claims that with the makes-sense task, subjects pay more attention to the formal properties that differentiate deep and surface anaphors. If this is true, however, it is unclear why judgments for deep anaphors, which should be perfectly acceptable in the absence of a recent linguistic antecedent, were negatively affected. If subjects were ‘paying attention’ to the formal properties of deep anaphors, then the decrease in positive judgments for deep anaphors is unexpected. It is equally unclear why judgment times increased only for surface anaphors. Under Murphy’s model, both deep and surface anaphors are resolved in the same way. Both were negatively affected in terms of the judgment data, but it is not clear why subjects would take longer to render the judgments for surface anaphors but not for deep. An analysis on the responses to positive versus negative responses might shed some light on this issue; perhaps, for example, subjects took longer to give positive responses for the surface anaphor than the deep. However, this is not provided.
2.4 Summary of experiments on the distance effect

I have reviewed five experiments that have investigated the effect of distance on deep and surface anaphors. The combined results of Murphy (1985a) Experiments 1 and 2 showed that both deep and surface anaphors were affected by distance from the antecedent. However, the crucial interaction of distance x anaphor type was not provided, so we do not know whether surface anaphors were more affected than deep. Also, we do not know what kinds of intervening sentences Murphy used in the distant condition. We have seen that this is an important aspect of the design of the materials in testing distance.

The results of Tanenhaus et al. (1985) indicated that surface anaphors were affected by distance, but deep anaphors were not. However, no statistics were provided and, more importantly, there was the confound of ambiguity for surface anaphors.

Murphy (1990) attempted to shed light on the conflicting results by investigating the effect of ambiguity and the effect of task (comprehension time versus makes-sense task). He found that ambiguity had no effect; however, we noted that his materials differed from those in Tanenhaus et al. in possibly critical ways. Murphy did find an effect of task. In the comprehension task, both deep and surface anaphors were affected by distance. In the makes-sense task, both were affected in terms of decreased positive judgments for distant antecedents. However, the latency data showed that only responses for surface anaphors were slowed in the distant condition.

Murphy suggests the comprehension task taps ‘normal’ reading and integration processes, whereas the makes-sense task forces subjects to consciously attend to the formal differences between anaphors. This intuitively makes sense, but the pattern of results for Murphy’s makes-sense experiment is not fully explained by the suggested difference.

The clearest results on the effect of distance in normal comprehension appear to be those of Murphy (1990) Experiment 1, which measured reading times, and found both deep and surface anaphors to be affected by distance. Moreover, in this experiment subjects were given true-false questions that sometimes referred to the anaphor, but sometimes did not. The advantage of this aspect of the design is that it reduces the possibility that subjects will develop unnatural strategies in reading the anaphor sentence and thus may provide a better picture of what people actually do during normal reading comprehension.

The one potential problem area lies in the nature of the intervening sentence. Other than a discussion of the ambiguity manipulation, Murphy does not provide any details regarding the intervening sentences used in the distant condition. It is possible that these sentences introduced topic shifts or shifts in focus that could have affected both deep and surface anaphors.

Another point worth noting is that there may also be a broader effect of the task on these results. If subjects reading these anaphors are consistently finding the antecedent in the linguistic context, then they might pay more attention to this context. In other words, it may not necessarily be the case that every deep anaphor searches the recent linguistic context for an antecedent before, or as well as, considering antecedents in the mental model or environment.

Overall, then, the effect of distance on deep and surface anaphors remains unclear. It may be that the slow-down observed for both types of anaphors in Murphy (1990)
Experiment 1 is due to both failing to find a linguistic antecedent and having to employ pragmatic reasoning. Alternatively, it is also possible that the slow-down is due to different factors or combinations of factors for the different anaphors.

3. Effects of parallelism

In the previous section I discussed experimental investigations into whether deep and surface anaphors are differentially affected by distance from the antecedent. The idea behind this research is that, since syntactic/surface information decays rapidly, an anaphor that requires that its antecedent be available in its surface representation will be more detrimentally affected by distance from that antecedent.

In this section I review several studies which have investigated the so-called ‘parallelism’ effect. These studies use the fact that surface anaphors require a kind of ‘consistency’ or parallelism of their antecedent, while deep anaphors do not. This is illustrated in the following example:

(14) The garbage needed to be taken out by someone

a. so Bill did it. (deep; do-it)
b. *so Bill did. (surface; VPE)

It is generally assumed that an ‘active’ ellipsis cannot take a passive VP as its antecedent, whereas ‘do-it’ can take both active and passive antecedents. Unlike with the distance effect, where intuitively deep anaphors are more acceptable than surface anaphors when separated by distance but are no more grammatical per se, here we have a difference in grammaticality. According to Hankamer & Sag (1976), Sag (1977), this difference in grammaticality is based on the fact that an active VP cannot delete under identity with a passive VP as the identity condition is not met at LF. However, this issue of the difference in grammaticality is not as clear cut as it may seem, and I return to this issue below.

3.1 Murphy (1985a), Experiment 3

The first study that I will discuss here is Murphy (1985a) who, in a third reading time experiment (the first two are discussed in the preceding section) investigated the effects of distance crossed with consistency/parallelism of antecedent and anaphor type.

The materials for this study consisted of texts with consistent antecedents (active sentences) and inconsistent antecedents (passivized versions of the active sentences, in some cases adjusting word order to achieve naturalness). The antecedent sentence was either the fifth or sixth line of the text, the former case corresponding to the distant condition and the latter corresponding to the close condition. The anaphor sentence was always the seventh line. The texts were followed by true-false questions which sometimes referred to the anaphor sentence but not always (this, Murphy claims, was to prevent subjects from developing unusual strategies when reading the anaphors). The following example is based on Murphy’s discussion; these lines would appear following four or five lines of text (unfortunately no examples of an entire item are provided):
(15) Leslie kicked the ball.

The ball was kicked by Leslie.

[Intervening sentence] (distant condition only)

Fran did too. (surface)
Fran did it too. (deep)

The results indicate a main effect of distance, with both surface and deep anaphors taking longer to read when the antecedent is remote (marginal by subjects, \( p < .10 \), significant by items \( p < .005 \)). There was no interaction of distance and anaphor, indicating again that surface anaphors were not more affected by distance than deep. However, we do not have any examples of his materials, and so do not know what kinds of intervening sentences were used in the distant condition. There was also no interaction of consistency with anaphor type, indicating that surface anaphors were no more affected than deep anaphors.

Inconsistent antecedents slowed reading times by 259 msecs in the close condition, but reading times sped up by 30 msecs in the distant condition. The interaction of distance and consistency was significant over subjects \( (p < .04) \) but not over items \( (p = .20) \). Further investigation into the lack of significance by items revealed that the effect of consistency obtained only in the second half of the experiment; consistency had virtually no effect in the first half \( (by subjects) \). Murphy suggests that this may be due to subjects adopting an initial strategy of accessing a discourse model to resolve all anaphors, and then switching to a ‘copying’ procedure, which accesses linguistic form. Whatever the explanation may be for the different results in the first versus the second half, there was no difference between deep and surface anaphors in either half.

It is worth noting, however, that a number of factors are relevant in manipulating consistency. Mauner, Tanenhaus & Carlson (1995) found that whether the passive sentences are long (have agent by-phrases) or short can affect results \( (see below) \). It is unclear whether Murphy’s materials contained long or short passives or a mixture of both. Nevertheless, the decreasing effect of parallelism with distance for both deep and surface anaphors is an interesting result, one that supports the idea that both are affected by the available linguistic context when close.

3.2 Tanenhaus & Carlson (1990)

Tanenhaus & Carlson (1990) conducted three experiments comparing deep and surface anaphors with antecedents which varied in parallelism. They discuss the results of Murphy 1985a, Experiment 3 \( (discussed above) \), which showed that consistency of the antecedent affects both deep and surface anaphors in the close condition, a result clearly incompatible with Sag & Hankamer’s (1984) hypothesis that they are processed differently.

In addition, Murphy’s results are inconsistent with results obtained by Tanenhaus & Carlson (1984), where they observed processing differences using a makes-sense judgment task.
T&C (1990) propose that the different results are due to the difference in task; they suggest that in reading time tasks, subjects are typically told to move on to the next frame once they have ‘understood’ the current frame, leaving what is meant by ‘understand’ up to the subjects. They claim that subjects are ‘often confused’ (p. 263) by what experimenters mean by ‘understand’ and end up setting their criterion for understanding based on the questions that follow. They claim that a virtue of their approach is that the makes-sense task allows subjects to set a criterion based on sentences that do not make sense.

However, in both cases the nature of the materials—i.e., either the questions or the sentences that do not make sense—is under the control of the experimenter. That is, it would appear that the potential for these to fail to engender the desired processing is virtually the same. In one case the experimenter needs to pay special attention to the kinds of questions being asked, and in another the experimenter needs to ensure that the nonsensical sentences are non-sensical in a way that encourages subjects to develop a general reading strategy that attends to those aspects of ‘sense’ that are relevant for the particular study. We have already discussed the possibility that questions that relate solely to the anaphoric sentence may induce more conscious processing, in effect mimicking the type of processing found in the makes-sense task. Tanenhaus & Carlson suggest that the makes-sense task forces subjects to take the antecedent into account. Again, the same thing could be accomplished with questions, but also there is the issue of exactly what aspects of the antecedent subjects are attending to: are they looking for formal/grammatical properties, or are they considering meaning, which is essentially what a query about ‘sense’ pertains to?

However, before this issue is discussed any further it might be useful at this point to review the details of the three experiments.

In Experiment 1, T&C vary whether the antecedent is in the active or passive voice. In this way, their experiment is similar to Murphy 1985a, Experiment 3. However, instead of using 7-line texts, T&C use 2-line items, with the antecedent in the first line, and the anaphor sentence in the second line. Note that T&C are not measuring the effect of distance. The following example is taken from their paper:

(16) Someone had to take out the garbage (active-parallel)
    The garbage had to be taken out. (passive-nonparallel)
    But Bill refused to do it. (deep anaphor)
    But Bill refused to. (surface anaphor)

Subjects were presented with these items and then asked to indicate whether or not the anaphor sentence made sense, given the preceding context sentence. Two examples of an instance where the second sentence did not make sense are provided below:

(17) After the exam Bill decided to have a beer or two.
    Sam didn’t either.

(18) Yesterday, the sports star announced his retirement.
    Sam denied it, too.
T&C do not specify any particular criteria that they followed in constructing the non-sensical materials, and do not provide any more examples in the list of materials at the end of the article, which is unfortunate given their importance in establishing a rejection criterion. In both examples, it does not appear to me that subjects would have to pay particular attention to the form of the antecedent. That is, in deciding whether the second sentence made sense, the criterion is not necessarily something that would encourage the kind of processing that would differentiate between deep and surface anaphors any more than follow-up questions would.

The results are reported for two dependent variables: judgment latencies on positive judgments (hereafter the latency data) and percentage of sentences judged to make sense (hereafter the judgment data).

Their results show an interaction between type of anaphor and parallelism for the judgment data: percentage of positive judgments decreased significantly for surface anaphors (89% in the parallel condition to 70% in the non-parallel condition) but not for deep anaphors (94% for parallel, 91% for non-parallel).

However, no such interaction is found with the latency data. Both deep and surface anaphors are affected by lack of parallelism in terms of response times for positive judgments. This pattern is essentially the same as Murphy’s reading time result which indicated that reading times for both deep and surface anaphors were affected by lack of parallelism (in the close condition).

T&C find a similar pattern of results with Experiment 2, which, instead of using active-passive pairs, used verbal-nominalized pairs. For example:

(19) Verbal (parallel): It would do you good to jog into town.
Nominal (non-parallel): A jog into town would do you good.
Surface: Please let me know if you decide to.
Deep: Please let me know if you decide to do it.

Whereas deep anaphors are grammatical following the nominal form, surface anaphors are not. Once again, the judgment data confirmed this intuition: percentage of positive judgments decreased significantly for surface anaphors (89% to 71%) but not for deep anaphors (86% for both). The latency data repeated the results of Experiment 1, showing that both deep and surface anaphors were affected by lack of parallelism.

In Experiment 3, T&C replace ‘do it’ with another deep anaphor (null complement) to rule out the possibility that phonological realization of the anaphor has an effect. So, in comparing ellipsis to ‘do it’, one might argue that the fact that the latter is phonologically realized makes a difference. Null complement anaphors represent a deep anaphor with no phonological features. In this study, they use both nominal and passive non-parallel antecedents (9 passive and 2 nominal). The following is an example:

(20) Someone has to take out the garbage.
The garbage has to be taken out.
But Bill refused to. (surface)
But Bill refused. (deep)
Again the results indicate that judgments are detrimentally affected for surface anaphors only. Percentage of positive judgments for surface anaphors decreased from 95% in the parallel condition to 77% in the non-parallel condition whereas deep anaphors decreased from 92% to 89% across the same comparison. Also echoing the results of the first two experiments is the latency data: response times for both deep and surface increased in the non-parallel condition.

Thus, in terms of the judgment data, it would appear that these results support H&S’s proposal that deep and surface anaphors are processed differently, but not in terms of the latencies.

A number of things bear mentioning at this point. It is no surprise that positive judgments for ellipses following a non-parallel antecedent decreased. Regardless of what strategy subjects were employing in deciding which sentences made sense, these sentences are ungrammatical (or at least less grammatical than the parallel counterparts - see discussion below on degrees of grammaticality), whereas the deep ones are not. It is expected, then, that acceptance ratings would be lower for ungrammatical sentences that nevertheless at some level ‘made sense’ (as demonstrated by the 70% acceptance rate). This difference, however, does not necessarily provide clear evidence for a processing difference. It could be that deep anaphors also accessed the representation that surface anaphors, by hypothesis, are supposed to access, but found in that representation a grammatical antecedent, accounting for the stable judgments.

Second, I would like to point out a potential problem with the materials that has been overlooked in this study, as well as in Mauner, Tanenhaus & Carlson (1995), discussed below. Consider the following:

(21) a. The garbage was taken out.
    Bill did.

b. The garbage was taken out by Jack.
    (because) Bill refused to.

c. The garbage was taken out by Jack
    (because) Bill didn’t/hadn’t.

d. The garbage needed to be taken out
    so Bill did.

e. The garbage needed to be taken out by someone
    so Bill did.

f. The garbage needed to be taken out
    but Bill refused to.

g. The garbage needed to be taken out by someone
    but Bill refused to.

h. The garbage needed to be taken out
    but Bill didn’t.

All of these involve a passive context sentence followed by an active surface anaphor. And yet, it seems to me that there is a considerable range in acceptability. The
first, for instance, seems quite bad. But some of the others seem close to fine, or only marginal. (21c) for instance, which contains an Agent by-phrase in the antecedent sentence, seems much better than (21a). Also, sentences with ‘infinitival’ ellipsis (21b,f,g) seem more acceptable with non-parallel antecedents generally than those with auxiliary ellipsis (21a,c,d,e,h). Further, the type of connective seems to have an effect as well, with subordinate conjunctions improving the acceptability compared to coordinating conjunctions (21d vs. 21h).

A later study by Mauner, Tanenhaus & Carlson (1995) controls for the presence of an Agent by-phrase, albeit not for the same reasons (discussed below), but does not control for different kinds of ellipsis (infinitival vs. auxiliary) or the connective.

It is likely, then, that had the materials been controlled for level of acceptability the judgment data would have been different (either increased or decreased percentage of positive judgments, depending on what type of structure was chosen). It is also possible that the latency data would have differed as well, with increased times associated with the more ungrammatical types. This could have resulted in the missing interaction in the latency data.

3.3 Mauner, Tanenhaus & Carlson (1995)

Mauner, Tanenhaus & Carlson (1995) report on two follow-up studies to T&C (1990). They observe that the passive sentences used by T&C contained a mix of long passives with Agent by-phrases and short passives without Agent by-phrases. They suggest that the long passives were somewhat awkward, and that this may have affected the results of T&C’s experiment. They conduct an item analysis and find that short passives seem to have exhibited a much stronger parallelism effect for VPE than long passives, and conduct two separate experiments to test this difference.

In the first experiment, MT&C use active sentences with ‘someone’ as subject and short, agentless passives, for example:

(22) Someone needs to feed the kitten. (active- parallel)
    The kitten needs to be fed. (passive- non-parallel)
    Joey forgot to again. (surface)
    Joey forgot to do it again. (deep)

Again they use the makes-sense judgment task used in T&C. The judgment data indicate that both type of anaphor and parallelism had a significant effect, and also that the interaction of these two factors was significant. Positive judgments for VPE decreased from 94% in the parallel condition to 82% in the non-parallel condition, whereas positive judgments for do-it remained steady at 96% across the same comparison. This replicates the judgment data from T&C’s Experiment 1 that showed that positive judgments for VPE decrease significantly with a non-parallel antecedent compared to do-it.

Where the two studies differ is in the latency data. Recall that T&C found that both deep and surface anaphors were adversely affected by lack of parallelism in the latency data. Here, however, MT&C obtain an interaction of parallelism with anaphor type; in other words, the judgment and latency data now fall together, with an overall finding that
surface anaphors are detrimentally affected by lack of parallelism in terms of both percentage of positive judgments and the time taken to indicate those judgments (2979 msecs in the parallel condition vs. 3531 msecs in the non-parallel), whereas deep anaphors are not affected in terms of either judgments or times (3075 msecs vs. 3149 msecs).

In their second experiment, MT&C use only long passives with Agent by-phrases; these by-phrases were always ‘by someone’ so that the interpretation would better match the active sentences and the agentless passives of the first experiment.

The results for the judgment data indicate that the only significant effect was parallelism - both deep and surface anaphors were judged to makes sense less often following a non-parallel antecedent.

The latency data showed the same result: only parallelism reached significance. Judgment times for both deep and surface anaphors were increased following non-parallel antecedents.

MT&C suggest that previous findings of parallelism effects for both deep and surface anaphors (Murphy 1985a, Experiment 3, and the latency data for Tanenhaus & Carlson, 1990, Experiment 1) might be explained by factors unrelated to syntactic parallelism; they offer the awkwardness of the long passives used in their study and in Tanenhaus & Carlson (1990) as a possible explanation for why deep anaphors might be affected in this condition. They offer the following as an example of an ‘awkward’ construction used in T&C (1990):

(23) The valuable antique vase which belonged to Mrs. Jones was broken by John.

They further suggest that the phrase ‘by someone’ (which was used as the Agent by-phrase in all of their long passives) is ‘awkward except in quite restricted contexts’ (1990:10). In effect, their claim is that whatever is causing the effect for deep anaphors, it is not structural/syntactic but some other factor. 7

They also note ‘one puzzling aspect of the data’ (1990:9): the parallelism effects for the surface anaphors in both judgments and judgment times were smaller than those found in their first experiment with short passives. Again their explanation is the awkwardness of the long passives used in their study.

The findings of the short-passive experiment are the expected result, given the grammaticality difference. That is, it is not surprising that positive judgments for surface anaphors decreased and moreover that it took more time to render these judgments for surface anaphors than for deep. Crucially, however, these results do not necessarily constitute evidence for a processing difference. It could be that deep anaphors also consult the recent linguistic representation, but find in that representation a grammatical antecedent. This would account for why deep anaphors are unaffected in terms of the judgment and latency data.

More puzzling is the fact that both were affected by non-parallelism in the long-passive condition. The fact that positive judgments decreased for deep anaphors is

---

7 Whether this explanation can be extended to Murphy’s (1985a) experiment on parallelism in unclear as the materials are not provided. However, it is worth noting that Murphy does indicate that adjustments were made in the passive sentences to improve naturalness.
unexpected given that deep anaphors are grammatical in these cases. This result is similar to Murphy’s (1985a) result with distance (Experiment 2)—positive judgments decreased for deep as well as surface anaphors with distance, even though intuitively only surface anaphors are awkward in these cases. In that case, however, only surface anaphors were affected in terms of latencies, whereas here, with parallelism, both are affected. It is possible that awkwardness was a factor—this would explain the effect on latencies. Moreover, it seems reasonable that awkwardness would have an effect regardless of task (comprehension time or makes-sense). However, this explanation still needs to be tested experimentally.8 I should point out that Mauner et al. do not attribute the awkwardness to the passive construction itself, only to the particular sentences used in their and T&C’s study; presumably, then, it would be possible to construct a set of ‘non-awkward’ materials for comparison.

3.4 Summary of parallelism

Murphy (1985a) Experiment 3 varied parallelism and distance in a comprehension time task. He found that reading times for both deep and surface anaphors were slower when the non-parallel (passive) antecedent was close compared to when it was far. He interpreted these findings as support for the idea that both deep and surface anaphors are affected by the recent linguistic context.

Subsequently, T&C (1990) conducted a similar experiment but with no distance variable (antecedents were always in the sentence immediately preceding the anaphor) and using a makes-sense task. They found that only surface anaphors were affected in the judgment data, but that both deep and surface anaphors were affected in the latency data.

Mauner et al. (1995) conducted two experiments, one using short passives and the other using long passives. They suggested that the awkwardness of the long passives might have resulted in increased latencies for deep anaphors in T&C (1990) and Murphy (1985a) Experiment 3, close condition. They found that only surface anaphors were affected (judgments and latencies) in the short passive condition, but both deep and surface were affected (judgments and latencies) in the long passive condition. They argue that previous findings showing deep anaphors to be affected by a lack of parallelism can be explained in terms of non-structural factors and thus do not provide evidence that deep anaphors access a linguistic level of representation.

8 An alternative explanation lies in a somewhat different take on the notion of parallelism. In Belangér (1999), a theoretical paper, I propose that the subject of a verb phrase ellipsis establishes a relation with the subject of the antecedent VP and ultimately shares that subject’s theta role. I noted that passives with agent by-phrases seem to be more acceptable non-parallel antecedents than agentless passives. It could be that with agentless passives, the subject of the ellipsis has no constituent with which to establish a link. With long passives, a constituent is present, but in a non-parallel position, accounting for the decrease in acceptability relative to an active antecedent. This might explain why Mauner et al. found that surface anaphors were less adversely affected by non-parallelism in the long passive condition relative to the short passive condition. While I did not extend my 1999 analysis to include ‘do it’, with deep anaphors, it could be that they are affected by the Agent constituent only if it is present. It could be that absence of an Agent has no effect, but that an Agent that is present but in a non-parallel position, has an adverse effect. At the moment, however, this is highly speculative.
However, the results of the short passive condition do not constitute clear evidence that deep anaphors do not access a linguistic level. The short passive provided a grammatical antecedent for deep anaphors, but not for surface anaphors. It is possible that a linguistic level was accessed with deep anaphors but that because the antecedent in that representation was grammatical, both positive judgments and latencies remained stable. If awkwardness is a factor in the long passive condition, then this would affect both types of anaphors. Thus, neither the short passive nor the long passive condition sheds any light on possible deep-surface processing differences.

Further, the results of Mauner et al. (1995) do not address the results of T&C (1990) Experiments 2 and 3, discussed above, which obtained differences between deep and surface in the judgment data but not the latency data.

In summary, the experiments that found both deep and surface anaphors to be affected by lack of parallelism have a possible alternative explanation in terms of awkwardness. That is, they do not provide clear evidence that both deep and surface anaphors access a linguistic level, although this possibility is by no means discounted. On the other hand, the one experiment that obtained a clear deep-surface difference in the results (Mauner et al. Experiment 1) does not provide evidence that different levels of representation are involved. This is because of the grammaticality difference between deep and surface anaphors under these conditions.

4. Particle Movement

I now turn to a discussion of TC&S’s (1985) experiment which we replicated and extended. Reporting on the results of this replication study is the main goal of this paper.

In their paper, TC&S report on various studies investigating levels of representation. They posit three ways in which anaphors/gaps might access their antecedents, each of these corresponding to different levels of representation: literal replacement, logical replacement and direct reference. Literal replacement involves accessing a surface representation to locate an antecedent. Investigation into this type would test for whether ‘surface’ properties of the antecedent (for example, phonological information) are active at the anaphor site.

Somewhat confusingly with respect to terminology, this is not the level they suggest is accessed with surface anaphors. Rather, they propose that the relevant level here is logical form (and hence involves logical replacement)–‘the anaphoric element is linked to the logical form (or meaning) of the linguistic antecedent rather than its surface form’ (p. 385).

Unlike literal replacement, logical replacement does not access phonological features, but it is unclear what other differences might be involved. Both of these are claimed to be linguistic levels—it seems as if they correspond to the theoretical levels of PF and LF, although TC&S do not state this explicitly. The current thinking in generative linguistic theory is that PF and LF are the only linguistic levels, and that there is no level of surface structure (or deep structure, for that matter) (Chomsky, 1995 and later). Phonological and word order information is encoded at PF; LF encodes semantically relevant information such as scope, binding and thematic relations. It is not clear to me how closely these correspond to what TC&S had in mind. It is also not clear how relevant the distinction is for our purposes, as we are not attempting at this point to probe the differences between
these two levels. Evidence that linguistic information of any kind (be it word order, phonological features, syntactic frame, etc.) is accessed with either a deep or surface anaphor we take to be evidence that a linguistic level is involved. Whether different linguistic levels are involved is another layer that we will not address here (but see the final section for some questions).

Finally, deep anaphors are proposed to involve direct reference, whereby the anaphor is ‘assigned the same denotation as its antecedent by being directly linked to some element in the constructed [non-linguistic] representation’ (p. 385).

They report on a number of studies, two of them comparing deep and surface anaphors. The first one is discussed above in Section 2.2 on the effect of distance.

The second experiment was intended to test in a more direct way whether deep and surface anaphors access different levels of representation during resolution.

In this experiment, subjects were presented with a sentence containing the antecedent, followed by a sentence containing either a deep or surface anaphor. This pair of sentences was followed by a verification sentence, which subjects had to rate as being either true or false. The dependent variable was the response time to the verification sentence.

Verification sentences varied from the antecedent sentence in one of two ways: either structurally or lexically. I will focus here on the structural changes, which were introduced by using an antecedent that contains a particle verb; the verification sentence either matched or mismatched the antecedent sentence in terms of the position of the particle and direct object. The following example is provided by the authors:

(24) Matching antecedent sentence: Jenny asked Ann’s boyfriend out yesterday.

Mismatching antecedent sentence: Jenny asked out Ann’s boyfriend yesterday.
Surface anaphor: Ann was furious that she did.
Deep anaphor: Ann was furious that she did it.
Verification sentence: Jenny asked Ann’s boyfriend out.

The idea is that only surface anaphors will access a level of representation where the relative order of the particle and direct object is encoded, and so keep that order ‘active’ in working memory. When a mismatching verification sentence is encountered after a surface anaphor, the different word order from the first sentence will still be ‘active’ and so will slow reaction times down compared to the matching condition (where the order that is active and the order in the verification sentence are the same). With deep anaphors, a non-linguistic level that is insensitive to word order is accessed. Therefore, response times to a mismatching verification sentence will not be affected, because the original order is no longer active, having undergone ‘normal’ decay.

An underlying assumption is that it is easier to process a verification sentence which has the same structure as the antecedent sentence than one that has a different structure. In other words, that there is some syntactic priming between like structures or inhibition between unlike structures. In this regard, Frazier, Taft, Roeper, Clifton & Ehrlich, (1984) found that, generally, the second clause of a co-ordinate construction was easier to process if it matched the first clause in structure. On the other hand, Branigan (1996) found that only locally ambiguous garden path sentences were subject to priming from the
previous structure. To my knowledge no one has specifically studied the extent to which a sentence primes itself — or alternatively, the extent to which a sentence inhibits processing of a second sentence to which it is almost, but not quite, identical (i.e., same lexical items and meaning but different structure).

Although no statistical tests are reported, the results show that response times for a mismatching verification sentence following a surface anaphor were slowed by 627 msecs compared to the matching condition (2,190 vs. 2,817), whereas response times actually decreased by 31 msecs following deep anaphors across the same comparison (2,320 vs. 2,289).

This result is quite striking, even in the absence of statistical tests. If reliable, it represents, to my mind, the single most compelling piece of evidence in favour of differential processing between deep and surface anaphors.

The main advantage of this design over previous experiments that modulate the particular form of the antecedent is that grammaticality is not a factor. In other words, both deep and surface anaphors are grammatical following either order of particle/direct object. Thus any slowdown that occurs after surface anaphors cannot be attributed to differences in grammaticality. Recall that this was a major problem in the experiments on parallelism. The advantage over length as a factor is that length may have also added complexity to the conceptual representation, thus muddying any differences between deep and surface anaphors. The advantage over using distance is mainly that it represents a more direct investigation into the kind of information encoded at the level of representation where resolution occurs.

TC&S interpret the results to show that surface anaphors access a linguistic level of representation to locate their antecedents, whereas deep anaphors access a non-linguistic level. They suggest that surface anaphors, by accessing the linguistic level of logical form, keep the linguistic structure (including word order) active, in a sense (temporarily) preventing it from undergoing the rapid decay that is associated with linguistic form. With deep anaphors, on the other hand, which access a non-linguistic level, the linguistic representation of the antecedent has undergone normal, ‘Sachs-type’ decay by the time the verification sentence is encountered.

4.1 A possible re-interpretation of TC&S

Although the results clearly indicate a difference between surface and deep anaphors, it is less clear what the precise nature of this difference is. It does seem evident that the word order of the antecedent VP is active after the surface anaphor but not after the deep. There are, however, a number of scenarios under which this difference might hold.

The first scenario, and the one put forward by TC&S, is that surface anaphors, in accessing a linguistic level during resolution, refresh or reactivate this level. Deep anaphors, on the other hand, have no effect on the linguistic representation of the antecedent and so this representation simply undergoes normal decay. Under this scenario, the word order representation is more active after the surface anaphor than it would otherwise be (i.e., under conditions of normal decay).

The second scenario is that surface anaphors do not in fact reactivate or refresh the linguistic representation of the antecedent at all, rather, it is the deep anaphor that acts on this representation in an inhibitory fashion.
The third and final scenario involves some combination of reactivation and suppression on the part of the surface and deep anaphors respectively.

The lack of a control condition to reveal the level of activation of the word order of the antecedent under normal decay conditions (that is, with no anaphor at all) is what makes it impossible to tell which of the above scenarios in fact holds. One might find, for example, that the level of word order activation following the surface anaphor is in fact no different than with no anaphor at all, suggesting that deep anaphors actively suppress this level during resolution. Alternatively, one might find that the word order activation following a no anaphor control is somewhere in between that for a deep and surface anaphor, indicating that both reactivation and suppression might be involved.

This idea of suppression with deep anaphors has not, to our knowledge, been put forth in the literature. However, Lucas, Tanenhaus & Carlson (1990) present evidence that resolution of NP-anaphors involves inhibition of inappropriate antecedents as opposed to reactivation of appropriate ones. There is therefore some precedent for the notion that the activation level of some entity (in their study a lexical item) can be suppressed through anaphor resolution. More relevant is the finding by Cloitre & Bever (1988) that repeated NP anaphors (which are argued to be surface anaphors) temporarily suppress the conceptual representation of the antecedent while accessing the surface representation of the same. We ask here whether deep anaphors might initiate a similar suppressive mechanism but of the surface rather than conceptual representation.

In the two experiments reported on below, we attempted to replicate and extend the results of the particle movement study, adding several conditions aimed at addressing the above questions.

5. Experiment #1

I have outlined TC&S’s particle movement experiment above. Although their experiment included lexical as well as structural manipulations, our studies investigated only the effects of structural manipulations. I should note that we were unable to obtain the original materials used in TC&S’s experiment, and so devised our own based on the description and example provided in their paper.9 An example is provided below:

9 Note that Carlson & Tanenhaus (1984) report on the same experiment but a different example is provided. The example they provide is as follows:

The old man didn’t want to throw out the old newspapers.
But when the health authorities threatened him, he did (it)

Matching verification: The old man threw out the old newspapers.
Mismatching verification: The old man threw the old newspapers out.

This example differs from the one quoted in the text in a number of ways. First, there is an ambiguity present for the surface anaphor, but not the deep. So, the surface anaphor can grammatically be interpreted as ‘He wanted to throw out the old newspapers’ whereas the deep anaphor cannot. Second, whereas the example in the text repeats the anaphor sentence almost verbatim (except for the exclusion of ‘yesterday’), the verification sentence differs quite significantly from the antecedent: it is simple in contrast to the antecedent sentence, which is complex, and whereas the antecedent VP is non-tensed, the verification VP is tensed. Although it is unclear what effect these factors might have had, it does not appear that the main finding is seriously undermined as both of them would have been present in both mismatch and match for
(26) **Matching antecedent:** The travel agent messed up Frank’s booking.

**Mismatching antecedent:** The travel agent messed Frank’s booking up.

**Surface anaphor:** He was very disappointed that she did.

**Deep anaphor:** He was very disappointed that she did it.

**Verification sentence:** The travel agent messed up Frank’s booking.

Mismatches were introduced by shifting the particle to the right of the direct object in the antecedent sentence, which was always a simple clause. The verification sentence was always in unshifted form (this aspect of the design was altered in the second version, where both antecedent and verification sentence had shifted and unshifted forms). The anaphor sentence always had a subject referring unambiguously to one of the referents introduced in the antecedent clause and a predicate adjective, with the anaphor contained in a subordinate clause.

We anticipated that we would replicate the results of TC&S, with response times to mismatching verification sentences being slowed more following surface anaphors than deep.

I argued above that the precise nature of the processing difference cannot be assessed without appropriate points of comparison. In the current study, we added three conditions. In one of these, the ‘Null Condition’, the verification sentence immediately followed the antecedent sentence. We expected this condition to reveal the maximum effect of mismatch because no material intervened between the two clauses (cf. Sachs’s 1967 zero-delay condition). This measure was intended to provide a point of comparison for the other conditions. In the event that surface anaphors reactivate the word order representation to original levels, we would expect that these two conditions would pattern similarly.

The second condition that we added was the ‘Adjective Condition’. In this condition, the intervening sentence was just the subject and predicate adjective from the surface and deep conditions. For example, the adjective condition for the above example would simply be:

(27) He was very disappointed.

This condition was the most speculative of the three additional conditions. We hoped that the mismatch effect observed here would give us some idea of the level of activation of the antecedent word order just prior to the introduction of the anaphor because it represents the fragment that appears just before the clause containing the anaphor. The main reason for our interest was the issue of reactivation, widely discussed in other literature on anaphora but not in the deep-surface debate. That is, other studies on, for example, pronouns or traces, have framed experiments on resolution in terms of reactivation (see, e.g., Love & Swinney, 1996; Fodor, 1993; McKoon & Ratcliff, 1994 among many others) - very loosely, a lexical item (antecedent) is encountered, its representation is activated, it subsequently deactivates and is reactivated again at the surface (and deep) anaphors, and a difference was obtained. Nevertheless, we avoided the ambiguity problem in our experiments and our verification sentences were verbatim repetitions of the antecedent sentence, modulo particle shift.
anaphor site (see General Discussion). We wondered how linguistic representations of VPs (as opposed to lexical items) might behave in this respect - does the antecedent become less active and then get reactivated at the anaphor?

However, we quickly noted a possible problem with this condition in terms of investigating this question. The problem was that intuitively these sentences involve a kind of ‘completion’, for example, *He was very disappointed (that the travel agent messed up his booking)*. It is unclear whether this reflects what actually happens in processing, i.e., whether readers/listeners treat the first sentence as a kind of antecedent; if so, one might expect a similarity with one of the other two anaphoric conditions. Despite the unknowns surrounding this condition, we nevertheless decided to include it, as we were curious as to how it would compare with the deep and surface conditions.

Finally, we added a ‘Neutral Condition’. In this condition, the intervening sentence, like the anaphor sentences, contained a pronoun referring back to the antecedent clause for coherence, but did not contain any other anaphors. The following is an example:

(28) He decided to spend the holidays at home.

We intended this condition to serve as the counterpoint to the Null Condition: whereas the Null Condition was expected to reveal the maximum effect of mismatch, the Neutral Condition was expected to reveal the level of activation of word order after a clause with neither ellipsis nor do-it—in other words, to reveal the ‘normal’ decay of word order information.

5.1 Method

Subjects: 60 university-educated (or university degree in progress) adult native speakers of English were paid for their participation.

Materials: 40 items were constructed as per the above criteria, 4 per condition. There were two factors: Match (2 levels) and Type (5 levels). Items were counterbalanced across 10 groups. 60 filler items and 8 practice items were also constructed. Items were separated into blocks of 25 which were presented in a random order and also randomized within blocks. The stimuli were presented using DMDX software programmed by Jonathan Forster at the University of Arizona.

Procedure: The subject initiated presentation of the first line (the antecedent sentence) of each item by pressing the Space Bar. Each line of the item (not including the verification sentence) consisted of a sentence centered on the screen that remained on the screen until the subject pressed the Space Bar to continue to the next line. After one or two lines (depending on the condition), the verification sentence appeared centered on the screen in capital letters. Subjects had 4 seconds to respond true or false to the verification sentence after which time the sentence disappeared from the screen; the subject would then have to press the Space bar to continue to the next item. True responses were provided by pressing the Right Shift key, false responses were provided by pressing the Left Shift Key. After the subject responded to the verification sentence, a blank screen appeared and remained until the subject pressed the Space Bar to continue to the next item.
Subjects were requested to sit with their hands on the keyboard with their thumb on the Space bar and fingers on each of the Shift keys in order to get the most accurate picture of the reaction times.

5.2 Results & Discussion

We employed the layered technique outlined in Van Selst & Jolicoeur (1994) to remove outliers, however, as it turned out, there were no outliers in this experiment, possibly due to the 4 second time limit on responses. The error rate for responses to the verification sentences was 8.5%. Incorrect responses were eliminated from the analyses.

We conducted a 2-way ANOVA with group as a between-subjects factor in the by-subjects analysis (Pollatsek & Well, 1995). The two within subjects factors were Match (2 levels: Match and Mismatch) and Type (5 levels: Null, Adjective, Surface, Deep and Neutral). Means are indicated in Table 2.

The results indicate a main effect of Match F1(1,59)=24.58, p<.00001, F2(1,39)=14.26, p<.001. Overall, verification sentences which matched the antecedent in word order were read more quickly than those which mismatched.

The effect of Type was marginally significant F1(4, 236)=2.38, p<.06; F2(4, 156)=2.19, p<.08. A Newman-Keuls test by items revealed that the verification sentences were read marginally faster in the Surface condition than in the Deep condition (p<.06). The Newman-Keuls test by subjects also indicated that verification sentences were read faster in the Surface condition than the Deep – this was marginally significant at p<.09 – and faster in the Surface condition than in the Null condition, also marginally significant at p<.08. The interaction of interest, Match by Type, was non-significant by both subjects and items: F1(4, 236)=1.77, p>.13; F2(4,156)=1.05, p>.38.

<table>
<thead>
<tr>
<th></th>
<th>Match</th>
<th>Mismatch</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>1613</td>
<td>1787</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>(370)</td>
<td>(436)</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td>1633</td>
<td>1664</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>(410)</td>
<td>(433)</td>
<td></td>
</tr>
<tr>
<td>Surface</td>
<td>1591</td>
<td>1669</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>(351)</td>
<td>(410)</td>
<td></td>
</tr>
<tr>
<td>Deep</td>
<td>1651</td>
<td>1756</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>(385)</td>
<td>(412)</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>1640</td>
<td>1702</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>(404)</td>
<td>(356)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Means in Milliseconds for Experiment Version 1. (Standard deviations in parentheses).

These results therefore failed to replicate the main finding of Tanenhaus et al.’s particle movement experiment. That is, while there was a main effect of Match, it did not affect responses to verification sentences more following surface anaphors than deep. In fact, the magnitude of the effect of mismatch (as indicated by the differences) was in the opposite direction than expected.
A number of potential problems were identified with this study that were investigated and, if necessary, corrected in Experiment 2, discussed below.

The first was that all of the verification sentences appeared in unshifted form. In the initial design of the experiment, an effort was made to use sentences that sounded natural in both the shifted and unshifted form; however, differences in markedness may have affected the results. If we suppose that the unshifted form represents the unmarked form, and the shifted one, the marked, then the mismatch condition would have involved reading the marked form in the antecedent sentence and the unmarked form in the verification sentence. The idea that the unshifted form might be less marked gains some support from Bock & Brewer (1974) who found that particle verb constructions exhibited better recall when unshifted. In terms of the effect of markedness on priming, it has been shown (Frazier, Taft, Roeper, Clifton & Ehrlich, 1984) that, relative to passive sentences, active sentences do not increase reading times of a following passive, but passive sentences (relative to active sentences) do increase reading times of a following active. Assuming that passive is, all things being equal, the marked case, it is relatively unaffected by the structure of the preceding sentence in terms of voice, whereas the active, unmarked sentence is affected. It is possible, then, that markedness played a role in our experiment, collapsing the differences across type.

Another problem was the nature of the Neutral condition. This condition was not controlled for number of clause boundaries; some sentences were simple clauses and others were complex, containing subordinate clauses. The sentences in the Surface and Deep conditions, on the other hand, were always complex. This may have resulted in an overall higher level of activation of the antecedent’s word order following the Neutral condition relative to the Surface and Deep conditions.

A third issue was the filler sentences, which did not contain any particle verb constructions where the answer to the verification sentence was false. This could have resulted in subjects noticing that all the particle verb constructions were followed by a true verification sentence, leading to an overall speeding up of reaction times. This was rectified in Experiment 2.

Noting the rather high error rate (8.5%), we thought that the time limit of four seconds to respond to the verification sentences might have been too short and might have negatively affected the depth of processing. We therefore increased the time limit to eight seconds.

Finally, the verification sentences in Experiment 1 were in capital letters to set them off from the rest of the sentences. However, it seemed that the capital letters were more difficult to read and moreover were quite visually different from the antecedent sentences. We therefore chose to represent verification sentences with the same font but underlined in Experiment 2.

6. Experiment #2

As noted above, this experiment differed from Experiment 1 in a number of respects. First, fillers included sentences with particle verbs where the answer to the verification sentence was false. This had been an oversight in the design of Experiment 1 and was corrected here.
Second, we varied both antecedent sentence and verification sentence in terms of whether the particle was shifted or unshifted. Therefore, instead of two factors, this experiment had three factors: Shift (2 levels: unshifted and shifted), Match (2 levels: match and mismatch), and Type (5 levels: Null, Adjective, Surface, Deep and Neutral).

We ensured that all sentences in the Neutral condition were bi-clausal because all sentences in the Surface and Deep conditions were bi-clausal.

Finally, to encourage subjects to read each item fully, we added yes-no questions to 10% of the items. This was because in the target items, all the verification sentences referred to the first line of the item. We wanted to ensure that subjects read the second line just as fully as the first. In addition to filler items where the verification sentence referred to the second line, we added yes-no questions that sometimes related to the second line (always for the target items).

6.1 Method

Subjects: 80 university students were paid for their participation.

Materials: There were three factors in this experiment: Shift (2 levels), Match (2 levels) and Type (5 levels). Materials consisted of 100 target items, counterbalanced across 20 groups. Each subject therefore saw five items per condition. 160 filler items were also constructed so that half of the total responses to the verification sentences would be positive and half would be negative. Filler items included 30 particle verb constructions. Twenty-six yes-no questions were also presented. Materials were presented using DMDX software.

Procedure: The subject initiated presentation of the first line (the antecedent sentence) of each item by pressing the Space Bar. Each line of the item (not including the verification sentence) consisted of a sentence centered on the screen that remained on the screen for eight seconds or until the subject pressed the Space Bar to continue to the next line. After one or two lines (depending on the condition), the verification sentence appeared underlined, centered on the screen. Subjects had 8 seconds to respond true or false to the verification sentence after which time the sentence disappeared from the screen; the subject would then have to press the Space bar to continue to the next item. True responses were provided by pressing the Right Shift key, false responses were provided by pressing the Left Shift Key. After the subject responded to the verification sentence, either a yes-no question or a blank screen appeared. The blank screen remained until the subject requested the next item using the Space Bar. The yes-no questions were in capital letters and centered on the screen. Subjects responded by using the shift keys, right for ‘yes’ and left for ‘no’. After the yes-no question, a blank screen appeared which remained until the subject pressed the Space Bar to continue to the next item.

Hand configuration instructions were given orally and were the same as for Experiment 1.
6.2 Results and Discussion

We analyzed the reading time data for the antecedent sentence (line 1) and the verification sentence separately. The results for the verification sentences are presented first.

6.2.1 Verification Sentence

Outliers were removed using the same method as for Experiment 1; outliers represented 3.2% of the responses. The average error rate for the verification sentences was 5%; for the follow-up questions the average error rate was 3%. Incorrect responses were removed. A 3-way ANOVA revealed a main effect of each of the factors. Means are in Table 3.

A 3-way ANOVA revealed a main effect of each of the factors. Means are in Table 3.

There was a Shift by Match interaction that was significant by subjects F1(1,79) = 5.92, p<.02; and by items F2(1,99) = 5.7, p<.02. A Newman Keuls test revealed that the basis of this interaction was that Shift had an effect within the Mismatch condition, but not the Match condition (means in Table 4).

<table>
<thead>
<tr>
<th></th>
<th>Unshifted</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Match</td>
<td>Mismatch</td>
<td>Match</td>
<td>Mismatch</td>
<td>Type</td>
</tr>
<tr>
<td>Null</td>
<td>1912</td>
<td>2105</td>
<td>1955</td>
<td>2162</td>
<td>2034</td>
</tr>
<tr>
<td></td>
<td>(547)</td>
<td>(724)</td>
<td>(623)</td>
<td>(722)</td>
<td>[2% error]</td>
</tr>
<tr>
<td>Adjective</td>
<td>1819</td>
<td>1929</td>
<td>1864</td>
<td>2021</td>
<td>1908</td>
</tr>
<tr>
<td></td>
<td>(563)</td>
<td>(541)</td>
<td>(584)</td>
<td>(673)</td>
<td>[2% error]</td>
</tr>
<tr>
<td>Surface</td>
<td>1789</td>
<td>1949</td>
<td>1802</td>
<td>2124</td>
<td>1916</td>
</tr>
<tr>
<td></td>
<td>(550)</td>
<td>(572)</td>
<td>(530)</td>
<td>(678)</td>
<td>[3% error]</td>
</tr>
<tr>
<td>Deep</td>
<td>1880</td>
<td>1903</td>
<td>1849</td>
<td>2014</td>
<td>1911</td>
</tr>
<tr>
<td></td>
<td>(636)</td>
<td>(548)</td>
<td>(573)</td>
<td>(684)</td>
<td>[4% error]</td>
</tr>
<tr>
<td>Neutral</td>
<td>1862</td>
<td>1989</td>
<td>1905</td>
<td>2049</td>
<td>1951</td>
</tr>
<tr>
<td></td>
<td>(566)</td>
<td>(633)</td>
<td>(566)</td>
<td>(663)</td>
<td>[4% error]</td>
</tr>
<tr>
<td>Match-Mismatch</td>
<td>1852</td>
<td>1975</td>
<td>1875</td>
<td>2074</td>
<td></td>
</tr>
<tr>
<td>Shift</td>
<td>1913</td>
<td></td>
<td></td>
<td></td>
<td>1975</td>
</tr>
</tbody>
</table>

Table 3: Means in Milliseconds for Experiment 2 (Standard Deviations by subjects in parentheses)
What this means is that when the verification sentence matched the antecedent sentence in word order, the relative position of the particle (shifted or unshifted) did not affect reading times. However, when the verification sentence mismatched the antecedent, the relative position of the particle made a difference. The means in Table 4 for the Mismatch condition show that reading times in the Shifted Condition are longer than for Unshifted. Recall that unshifted and shifted refer to the order of the antecedent, not the verification sentence. Thus, subjects took longer to read an unshifted verification sentence that followed a shifted antecedent than to read a shifted verification sentence that followed an unshifted antecedent. In other words, going from marked to unmarked is harder than the reverse. This is not unexpected given the findings with passive and active pairs in Frazier et al. (1984), discussed above.

The Match x Type interaction, the primary interaction of interest, was significant by subjects \( F(4, 316) = 2.7, p < .04 \) but missed significance in the by items analysis \( F(4, 396) = 2.08, p < .09 \). Means for this interaction are in Table 5.

<table>
<thead>
<tr>
<th>Match</th>
<th>Mismatch</th>
<th>Difference</th>
<th>p-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>1934</td>
<td>2134</td>
<td>200</td>
</tr>
<tr>
<td>Adjective</td>
<td>1841</td>
<td>1975</td>
<td>133</td>
</tr>
<tr>
<td>Surface</td>
<td>1796</td>
<td>2037</td>
<td>241</td>
</tr>
<tr>
<td>Deep</td>
<td>1865</td>
<td>1959</td>
<td>94</td>
</tr>
<tr>
<td>Neutral</td>
<td>1884</td>
<td>2019</td>
<td>135</td>
</tr>
</tbody>
</table>

Table 5: Means in Milliseconds for Match by Type

A Newman-Keuls test by subjects revealed that the interaction was based on the fact that all types were affected by mismatch except for the deep anaphor. This is a very interesting result, as it indicates that deep anaphors may involve suppression of the level of representation encoding word order. This is because verification reading times were affected for the control Neutral condition, indicating that, with a ‘normal’ decaying of structural information, word order is still active after an intervening bi-clausal structure. The fact that verification times were not affected by the mismatch with deep anaphors suggests that less word order information was present than would be available if the deep anaphor were not present.

This finding suggests that some kind of suppression is at work with deep anaphors, however, we also wanted to know whether surface anaphors reactivate or refresh the linguistic representation. If so, then we would expect the Surface condition to show more of an effect of mismatch than the Neutral condition.

We therefore conducted an analysis of the differences within each type between match and mismatch (see Table 6). There was a significant effect of Shift \( F_1(1, 79) = 6.11, p < .02; F_2 (1, 99) = 5.74, p < .02 \). There was also a significant effect of Type in the by-
A Newman-Keuls test on the Type effect found in the subjects analysis revealed that the two mean differences that differed significantly from each other were Surface and Deep anaphors (p<.03). None of the other conditions differed significantly from each other.

<table>
<thead>
<tr>
<th>Type</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>200</td>
</tr>
<tr>
<td>Adjective</td>
<td>133</td>
</tr>
<tr>
<td>Surface</td>
<td>241</td>
</tr>
<tr>
<td>Deep</td>
<td>94</td>
</tr>
<tr>
<td>Neutral</td>
<td>135</td>
</tr>
</tbody>
</table>

Table 6: Mean Difference Between Match and Mismatch

Taken together, a somewhat puzzling picture emerges. On the one hand, the Deep anaphor condition was the only condition that showed no effect of Mismatch, suggesting that, in comparison with the Neutral (no anaphor) condition, the word order of the antecedent is suppressed. On the other hand, the difference in the Mismatch effect between the Deep anaphor condition and the Neutral condition is not significant. Although Deep and Surface differ from each other, they do not differ from any other condition. Thus it appears that the Differences data cannot shed light on the question of reactivation with Surface anaphors or indeed any further light on the question of suppression. In other words, at least with respect to the Differences data, our control conditions, where the data lies somewhere in between Deep and Surface but not significantly different from either, do not illuminate the picture any further. Surface does not pattern with any control conditions that Deep does not also pattern with, and yet they are different in some way. The task now is to find control conditions that reveal either a three-way difference among Surface, Deep and Control (which would suggest both suppression and reactivation) or a two-way difference (which would suggest either reactivation or suppression). Given the results of the primary analysis, however, I will–tentatively–conclude that there is some evidence for suppression with Deep anaphors. At this point, we have no evidence either way regarding reactivation with Surface anaphors.

6.2.2 Line 1

The purpose of analyzing the data from Line 1 was to get some idea of reading times when the clause is encountered for the first time. The experiment is based on the assumption that the availability of the structural representation of the clause, which includes word order, will affect the reading time of the verification sentence, in other words, that some kind of syntactic priming is at work. A baseline reading time is therefore of interest.

The first analysis was an overall comparison between Line 1 and the Verification Sentence. This revealed that Line 1 is read more slowly than the verification sentence (2461 vs. 1958; F1 (1,79)=83.57, p<.0001; F2(1,99)=402.73, p<.00001). Thus there is an
overall effect of priming, which would include priming at the lexical and semantic levels as well as the structural level. It is easier to read a sentence that has recently been processed, even given that subjects had to make a true-false judgment on the verification sentences.

We also compared reading times for Line 1 as a function of Shift. As Match and Type could not have affected the reading times for Line 1 (as these factors came into play after Line 1), we calculated the means for Line 1 unshifted and Line 1 shifted and then conducted a two-tailed t-test for correlated samples. The means were not significantly different by either subjects or items (2457 for unshifted vs. 2465 for shifted, \( t_{79} = -0.35 \), \( p > .7 \) by subjects; \( t_{99} = -0.38 \), \( p > .7 \) by items). This result is interesting because we did obtain an effect of Shift in the verification sentences - shifted sentences were overall read more slowly. We also obtained an interaction with Match indicating that it is harder to go from Shifted to Unshifted than it is to go from Unshifted to Shifted. It is possible that even though there is no difference in the initial reading times for Shifted and Unshifted constructions, the shifted word order remains more salient, causing more interference when an unshifted word order follows. In general then, we might expect to find that marked syntactic frames have a somewhat different lifespan than their unmarked counterparts.

We also conducted an analysis of the differences between Line 1 and the verification sentence. We anticipated that this analysis would provide a picture of the extent of the priming effect across conditions. The magnitude of the mean difference between Line 1 and the verification sentence is interpreted as the magnitude of the overall priming effect, combined with any inhibition effects and the time taken by the judgment itself. Mean differences are indicated in Table 7.

<table>
<thead>
<tr>
<th>Match</th>
<th>Unshifted</th>
<th>Shifted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>490</td>
<td>471</td>
</tr>
<tr>
<td></td>
<td>323</td>
<td>217</td>
</tr>
<tr>
<td>Adjective</td>
<td>572</td>
<td>524</td>
</tr>
<tr>
<td></td>
<td>524</td>
<td>434</td>
</tr>
<tr>
<td>Surface</td>
<td>611</td>
<td>616</td>
</tr>
<tr>
<td></td>
<td>565</td>
<td>257</td>
</tr>
<tr>
<td>Deep</td>
<td>513</td>
<td>487</td>
</tr>
<tr>
<td></td>
<td>565</td>
<td>387</td>
</tr>
<tr>
<td>Neutral</td>
<td>602</td>
<td>694</td>
</tr>
<tr>
<td></td>
<td>565</td>
<td>348</td>
</tr>
</tbody>
</table>

Table 7: Line 1-Verification Line, Mean Differences in milliseconds

We found no effect of Shift (F1(1, 79)=2.74, \( p > .10 \); F2(1,99)=2.13, \( p > .14 \)) indicating that the overall extent of priming did not differ depending on whether Line 1 was shifted or unshifted. We found a main effect of Match (F1(1,79) = 30.71, \( p < .00001 \); F2(1,99) =35.54, \( p < .0000001 \)), indicating that there is a significantly greater extent of priming for the Match condition compared to the Mismatch (595 vs. 411). We also obtained a Shift by Match interaction (F1 (1, 79)=16.93, \( p < .0002 \); F2 (1,99)=12.52, \( p < .001 \)). The means are indicated in Table 8:

<table>
<thead>
<tr>
<th>Match</th>
<th>Unshifted</th>
<th>Shifted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>557</td>
<td>634</td>
</tr>
<tr>
<td>Mismatch</td>
<td>492</td>
<td>329</td>
</tr>
</tbody>
</table>

Table 8: Means in Milliseconds for Shift by Match: Line 1-Verification line
The interaction is based on the fact that within Unshifted, Match does not have a significant effect (p>.12 by subjects; p>.22 by items) whereas within Shifted, Match has a highly significant effect (p<.0002 by subjects and items). Shifted-Match at 634 has the most robust priming effect, again supporting our hypothesis that the shifted structure remains more salient (significant compared to Shifted-Mismatch p<.0002; Unshifted-Mismatch, p<.0004; and marginally significant by subjects compared to Unshifted-Match - p<.08 by subjects/p>.1 by items). Also confirming this is the fact that Shifted-Mismatch displayed the weakest priming effect, significant across all comparisons (Unshifted-Match p<.0002, Unshifted-Mismatch, p<.0004). This too would be expected if the shifted word order is more salient thus causing more interference when the unshifted word order is encountered.

There was also an effect of Type (F1, (4, 316)=4.2, p<.003; F2 (4, 396)=4.86, p<.001). Means by Type are displayed in Table 9:

<table>
<thead>
<tr>
<th>Type</th>
<th>Difference Line 1-Verification Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>375</td>
</tr>
<tr>
<td>Adjective</td>
<td>536</td>
</tr>
<tr>
<td>Surface</td>
<td>539</td>
</tr>
<tr>
<td>Deep</td>
<td>513</td>
</tr>
<tr>
<td>Neutral</td>
<td>552</td>
</tr>
</tbody>
</table>

Table 9: Means in Milliseconds for Type: Line 1-Verification Line

These results indicate that the Null condition displayed the weakest priming effect. This was significant compared to all other types (all p’s<.01). None of the other types differed from each other. This was somewhat surprising given that one would expect the structure to be very salient immediately following the sentence. Again, however, it is possible that a repetition effect attenuated the priming effect.

We did not obtain a Match x Type interaction (F1 (4, 316)=.52, p>.7; F2 (4, 396)=.66, p>.6). In other words, deep and surface anaphors did not have different effects on the extent of priming. Both exhibited higher degrees of priming in the Match condition than in the Mismatch. Means are shown in Table 10.

<table>
<thead>
<tr>
<th>Type</th>
<th>Match</th>
<th>Mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>476</td>
<td>269</td>
</tr>
<tr>
<td>Adjective</td>
<td>598</td>
<td>481</td>
</tr>
<tr>
<td>Surface</td>
<td>674</td>
<td>406</td>
</tr>
<tr>
<td>Deep</td>
<td>588</td>
<td>435</td>
</tr>
<tr>
<td>Neutral</td>
<td>633</td>
<td>450</td>
</tr>
</tbody>
</table>

Table 10: Means in Milliseconds of Line1-Verification Time; Match by Type

It is unclear how this finding impacts on the interpretation of the results of the verification sentence alone. If it is true that Deep anaphors suppress the word order information, then there should be no facilitative priming occurring with the Match Condition relative to the Mismatch Condition. At present I have no explanation to offer
other than to say that the finding of suppression should be interpreted with some caution at this point.

7. General Discussion

The psycholinguistic literature comparing deep and surface anaphors has yielded varied and often conflicting results. I have conducted two replication and extension experiments based on the particle movement experiment outlined in Tanenhaus, Carlson, & Seidenberg (1985). The results of their experiment indicated a clear difference in processing between deep and surface anaphors, but lacked the appropriate control conditions. Our Experiment 2 replicated their main finding in the subjects analysis. Reading times for verification sentences following surface anaphors were significantly slowed in the Mismatch condition, whereas reading times following deep anaphors were unaffected. Thus the claim that deep and surface anaphors are resolved differently was supported.

A number of interesting additional observations came out of the present study, the main one being suppression with deep anaphors. This was based on the fact that the Deep condition showed no effect of Mismatch while the Neutral, no-anaphor condition did.

We also found interesting results with respect to markedness and the shifted order of verb particle constructions. Our results indicate that the shifted word order had a greater effect on the processing of the verification sentences (both those that matched and those that mismatched) than the unshifted word order, suggesting that its syntactic frame is somehow more resilient and/or salient.

Another finding worth noting is the priming effect: overall, matched verification sentences were easier to process than mismatched. Recall that Branigan (1996) found that priming in comprehension (as opposed to production) is a competitive effect and only arises with locally ambiguous garden path sentences where two or more structures compete at a certain point in processing. In other words, processing one garden path sentence eases comprehension of a following garden path sentence, but processing of, for instance, a double-object construction does not make processing a following double-object construction any easier than a prepositional dative. This is because there is no point at which two possible structures compete. Our results, however, indicate there is more to priming in comprehension than a competitive effect, as particle constructions involve no choice that needs to be made by the parser and yet we found evidence of priming.

In considering the question of (re)activation and suppression, it is instructive to consider other experiments on anaphoric resolution, many of which have focused on anaphors reactivating their antecedents. Reactivation occurs when an antecedent undergoes initial activation followed by deactivation followed by reactivation at an anaphor site. To show reactivation, one must of course show that the antecedent’s activation level increases at or after the anaphor relative to before the anaphor.10 Two methods have been used to investigate the question of reactivation: cross-modal lexical decision or naming and visual probe recognition (note that these only test the activation levels of a lexical item, not of a phrase). These have been used to investigate definite NP

10 Recall that this was our original reason for including the Adjective condition.
anaphors (Dell, McKoon & Ratcliff, 1983) reflexives (Nicol & Swinney, 1989), pronouns, wh-trace (Love & Swinney, 1996), NP-trace (Osterhout & Swinney, 1993), and PRO (Bever & McElree, 1988), with varying results. Fodor (1993) reports that visual probe recognition tasks show reactivation for all anaphor types but generally only at the end of the sentence, whereas cross-modal priming experiments show immediate reactivation effects for wh-traces, reflexives, and pronouns, but not NP-traces or PRO.

Fodor suggests that the cross-modal task is sensitive to a linguistic level of representation\(^{11}\) whereas the probe recognition task is sensitive to a semantic level.

Note that attention has generally been restricted to investigating the activation/reactivation of lexical items, not of word order representations, as is relevant for surface anaphors whose antecedents are phrasal. In terms of the current study, we had hoped to probe whether surface anaphors induce a refreshing of the structure of the antecedent, in a similar fashion to the reactivation of lexical items found with other anaphors. We found no evidence for this, however, it should be stressed that the control conditions used in this study—in particular, the Neutral condition—may need to be rethought. The key is to construct a Control condition that will reveal whether the Deep-Surface difference is based on suppression or reactivation or both. Since the Neutral condition did not differ from either in the Differences data, these questions remain.

The second issue that is relevant with respect to the above mentioned studies on reactivation is that, if we take Fodor’s proposed distinction between the two tasks (cross-modal priming and probe recognition) seriously, then these experiments provide evidence that a linguistic level is consulted with deep anaphors. If the cross-modal task is sensitive to a linguistic level and pronouns (a deep anaphor) show immediate reactivation effects with this task, then that suggests that a linguistic level is implicated in the resolution of deep anaphors. There is further evidence to support this claim. Carreiras, Garnham, & Oakhill (1996) report on a Spanish study that found that grammatical gender features of the antecedent (which have no semantic correlate) had an effect on resolution of pronouns, thus providing evidence that purely linguistic information is accessed. Cloitre & Bever (1988), using a naming task, suggest, based on their findings, that definite NP anaphors (another deep anaphor) access a surface level of representation that in fact encodes phonological information (but see Lucas, Tanenhaus & Carlson, 1990, who obtained a different result). Thus there is evidence that deep anaphors are affected by purely linguistic factors which presumably would not be encoded in the mental model.

Further evidence that deep anaphors are affected by linguistic representations comes from work on parallel function (see, for example, Sheldon, 1974; Smyth, 1994 and others). Resolution of pronouns is affected by the grammatical and semantic role of the antecedent, with faster resolution times associated with antecedents that have the same grammatical and thematic roles as the pronoun and which occur in structurally congruent clauses.

---

\(^{11}\) Initially she identifies this level as s-structure, but later considers and explores the possibility that the relevant level is PF. It is unclear how this would affect resolution of surface anaphors — certainly it goes contrary to Tanenhaus et al.’s position that the level of representation involved in surface anaphors is not affected by phonological features.
The results of the current experiment indicate that deep and surface anaphors are processed differently. However, the above studies indicate that at least some kinds of linguistic information are relevant for deep anaphors.

One way of approaching this is to maintain that there are two distinct levels, one linguistic and the other conceptual but that deep anaphors are in fact affected by the former, at least in some ways or under some conditions. So it may be that something like whether a particle is shifted or not is not “relevant” (and therefore may not disrupt processing), but that the voice of a construction is (a possibility given the mixed results reviewed above). In other words, there may be some linguistic properties that are more salient than others, and these might affect both deep and surface anaphors.

Alternatively there may be some level that mediates between the surface and conceptual representations, a level which encodes some linguistic information but not all (see Garnham, Oakhill, Ehrlich & Carreiras, 1995, for some discussion), and it is this level which deep anaphors tap during resolution.

Hopefully further research will provide insight into these questions.

References


Nicol, J. (1993) 'Reconsidering reactivation.' In G.T.M. Altmann & R. Shillcuck (eds.)
*Cognitive models of speech processing.* Hillsdale, NJ: Erlbaum.
sentence comprehension.’ *Journal of Psycholinguistic Research* 18:1, 5-19.
Nicol, J., J.D. Fodor, D. Swinney (1994) ‘Using cross-modal lexical decision tasks to
investigate sentence processing.’ *Journal of Experimental Psychology: Learning,
Memory and Cognition* 20:5, 1229-1238.
comprehension of verbal passives.” *Journal of Psycholinguistic Research* 22:2, 273-
286.
research: A suggestion for a better and more powerful analysis.’ *Journal of
Experimental Psychology: Learning, Memory & Cognition* 3, 785-794.
Sachs, J. (1967) ‘Recognition memory for syntactic and semantic aspects of connected
discourse.’ *Perception and Psychophysics* 2:9, 437-442.
2:1A, 95-100.
Philosophy* 7, 325-345.
reconstruction and semantic influence.’ *Journal of Psycholinguistic Research* 24:6,
517-532.
Sheldon, A. (1974) ‘The role of parallel function in the acquisition of relative clauses in
English.’ *Journal of Verbal Learning and Verbal Behavior* 13, 272-281.
between short and long term memory during anaphor comprehension.’
of Psycholinguistic Research* 23, 197-229.
*Proceedings of the 14th Annual Meetings of the Northeastern Linguistic Society.*
Amherst, Mass: GLSA.
representations?’ In D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.), *Natural
language parsing: theoretical, psychological and computational perspectives.*
Cambridge: Cambridge University Press.
of the 8th Annual Meeting of the Cognitive Science Society.* Lawrence Erlbaum