AMBIGUITY AVOIDANCE IN ENGLISH RELATIVE CLAUSES

DAVID TEMPERLEY

Eastman School of Music, University of Rochester

Are syntactic choices influenced by the need to avoid ambiguity? Studies of the use of *that* with English embedded clauses have reached negative conclusions on this point. It is argued here that these conclusions may be premature. Statistical analysis of another phenomenon of English—use of the optional relative pronoun or complementizer with object relative clauses—in written text suggests that both AMBIGUITY AVOIDANCE and ANAPHORICITY contribute to syntactic choices. Ambiguity avoidance is shown to operate at a 'strategic' level, influenced by general considerations of syntactic structure, but not by lexical distinctions or pragmatic factors.*

1. AMBIGUITY AVOIDANCE. By all accounts, linguistic communication involves a syntactic representation of some kind, transmitted from producer to perceiver via a sequence of words. For this process to occur successfully, the syntactic structure must be inferable from the words: that is to say, ambiguity must be avoided. It is natural to wonder if this principle might have explanatory value in the study of syntax. Perhaps certain facts about syntactic structure—not all of which have been satisfactorily accounted for by other means—could be explained in terms of the need to avoid, or at least to minimize, ambiguity, and of the tendency of languages to develop in such a way that this need is met.

The idea of AMBIGUITY AVOIDANCE has received only occasional attention in studies of syntax. One important discussion is found in Bever 1970 (see also Bever & Langendoen 1971). Bever suggested that the requirement of a relative pronoun in English subject relative clauses (1a,b) may be due to the fact that such clauses would often be ambiguous, seeming like main clauses, without one.[1] In object relative clauses, where the relative pronoun is optional, this ambiguity does not arise (1c,d).

(1) a. *The man hired me was very tall.
b. The man who hired me was very tall.
c. The man I hired was very tall.
d. The man who(m) I hired was very tall.

Similar reasoning, Bever suggested, might account for the fact that a complementizer is required preceding a clause in subject position (2a,b); without one, there is again a danger of interpreting it as the main clause.

(2) a. *John is sick is quite evident.
b. That John is sick is quite evident.

We should note that the potential ambiguity arising from examples 1a and 2a—assuming that they were syntactically allowed—would only be temporary; the correct analysis would eventually be clear (at least in these cases). What is at issue, then, is not the ambiguity of the complete sentence, but a local ambiguity or 'garden-path' effect that might temporarily impede processing.

In a similar vein, Hankamer (1973) proposes that certain constructions may be prohibited on the grounds that they would routinely cause ambiguities. Hankamer focuses particularly on gapped constructions. For example, you cannot say *Max gave Sally a nickel, and Harvey a dime*, meaning that Harvey gave Sally a dime; the reason, accord-

* Thanks are due to Tom Wasow and an anonymous referee for many helpful suggestions on an earlier draft of this article, and to Paul von Hippel for statistical advice.

[1] Not all dialects of English require a relative pronoun in subject relative clauses (see n. 22).

ing to Hankamer, is that this sentence is open to another interpretation (that Max gave Harvey a dime). These observations by Bever and Hankamer, as well as several others, are brought together by Frazier (1985), who proposes what she calls the IMPERMISSIBLE AMBIGUITY CONSTRAINT: constructions that would lead to ambiguities or misanalyses on every occurrence tend to be prohibited.

Ambiguity avoidance has also been occasionally mentioned in connection with syntactic change. Sometimes, a change in a language appears to result in the loss of some kind of syntactic information, while another (more or less contemporaneous) change seems to introduce the same information in another form. For example, in Middle English, the loss of word endings indicating case roughly coincided with the development of fixed S-V-O word order, so that case was indicated by word order instead (Pyles 1971, Danchev 1991, Smith 1996). In such situations, it is tempting to suppose that the information-losing change (loss of word endings) created a problem of ambiguity which had to be resolved by other means (fixed word order)—though it may also be that the new form of the information was added first, making the old form redundant. Syntactic changes may also be triggered by phonological changes; Harris (1978) suggests that the loss of word-final /s/ in French caused an ambiguity between singular and plural noun forms, which was then compensated for by the addition of an obligatory determiner (*les*) on plural forms.[2]

Before proceeding, one other approach to syntax deserves some discussion. This is what might be called the 'processing' approach, as it attempts to explain syntactic regularities from the perspective of perception: what kinds of syntactic structures can be processed most rapidly and easily? An example is Kuno's 1974 study of 'center-embedded' constructions (like *The mouse the cat the dog chased chased ran*); according to Kuno, the difficulty of processing such sentences—and their rarity in language use—is due to the fact that they require several incomplete constituents to be maintained simultaneously. A more recent example is Hawkins's EARLY IMMEDIATE CONSTITUENT (EIC) theory (1994), which attempts to account for a range of syntactic phenomena in various languages. One phenomenon in English concerns verb phrases containing both an NP and a PP; while the normal order of consitutents in this case is VP NP PP, if the NP is much longer than the PP it tends to be shifted to the end.

(3) a. He [$_{VP}$ sold [$_{NP}$ the ring] [$_{PP}$ for five dollars]]
b. He [$_{VP}$ sold [$_{PP}$ for five dollars] [$_{NP}$ the gold ring that his grandmother had given him as a present when he was five years old]]
c. ?He [$_{VP}$ sold [$_{NP}$ the gold ring that his grandmother had given him as a present when he was five years old] [$_{PP}$ for five dollars]]

A similar phenomenon arises with two-word transitive verbs: if the object noun phrase is long, it will tend to be shifted after the particle. (One might say *I picked the book up*, but not *I picked the book that you wanted to read this weekend up*.) Again, these preferences for certain syntactic constructions over others are explained in terms of their ease of processing. Hawkins's EIC theory states that processing is facilitated if the daughter constituents of a node can be identified rapidly, within a short 'window' of a few words. In the case of verbs with NP and PP, if the NP is long, the daughter constituents of the VP (the V, NP, and PP) can be identified more quickly if the PP

[2] A related question is whether phonological changes themselves can be seen as functional—driven by considerations of meaning. For example, in cases where final consonants are deleted, does this occur more often in cases where the deleted consonant is informationally redundant? See Kiparsky 1982 for an argument in favor of this idea, Labov 1994 and Guy 1997 for more skeptical views.

is placed first. (This can be seen intuitively, as example 3b is much easier to process than 3c.)

While the processing approach offers convincing explanations of many phenomena, it is important to distinguish it from the ambiguity-avoidance approach. Facilitation of processing and avoidance of ambiguity are certainly related; as noted with regard to the Bever examples, ambiguous sentences are usually difficult to process. But the converse is not necessarily true; sentences like *The mouse the cat the dog chased chased ran* are certainly difficult to process, but is there at any point a sense of ambiguity or 'garden path'—a danger of entertaining an INCORRECT analysis? The primary issue is not ambiguity per se, but rather memory load: some structures impose greater memory demands than others. (According to Hawkins's EIC theory, the complexity of a sentence is related to 'the number of phrasal nodes whose structure must be computed simultaneously' [1994:60].) In short, facilitation of processing and ambiguity avoidance are not the same thing; we might regard ambiguity avoidance as an ASPECT of processing facilitation, along with other aspects such as those discussed by Kuno and Hawkins.[3]

2. AMBIGUITY AVOIDANCE IN SYNTAX: THREE POSSIBILITIES. If ambiguity avoidance is a factor in syntax, how might we expect it to manifest itself? There are three broad possibilities.

(i) Ambiguity avoidance could play a role in the formation of general syntactic principles—either in universal principles of syntax, or in rules of particular languages. What Bever proposes seems to be along these lines (the rules at issue here are language-specific rather than universal): the syntactic rule that relative pronouns are required in subject relative clauses but not object ones, Bever suggests, is due to the need to reduce ambiguity. Changes in syntactic rules driven by ambiguity avoidance would also fall into this category.

(ii) Ambiguity avoidance could influence syntactic strategies. By 'syntactic strategy', I refer to regularities of syntactic behavior that cannot be considered actual rules; often these arise in cases where there is more than one syntactically correct way to express something. An example discussed in §1 is the fact that where a VP has an NP plus another complement, the NP is more likely to be placed last if it is longer (though either order is syntactically correct). I cite this simply as an example of a well-documented syntactic strategy, not necessarily as a strategy that is motivated by ambiguity avoidance; rather, it appears to be due to other factors (Arnold et al. 2000). But it is certainly possible IN PRINCIPLE that there would be syntactic strategies motivated by ambiguity avoidance. (Other possible examples of this are considered in what follows.)

Strategic ambiguity avoidance might take the form of rather general strategies—syntactic patterns or constructions that are preferred or avoided. It might also involve more specific lexical factors: that is, the choice between two syntactic constructions might be affected by the specific words involved. This might be motivated by the fact that certain words within a syntactic category may be more versatile than others, and thus more prone to ambiguity. Again, examples of such 'lexical strategies' are presented below.

[3] Indeed, Hawkins himself adopts an ambiguity-avoidance explanation for certain phenomena. He observes, for example, that many predominantly head-final languages have head-initial positioning for complementizers of embedded clauses and for antecedent NPs of relative clauses. This may be due to ambiguity avoidance; placing the head of a dependent clause first prevents elements of the clause from being misidentified as main-clause elements.

(iii) Ambiguity avoidance could come into play in a highly situation-specific, ad hoc fashion. Whether a sentence is actually ambiguous may depend not only on broad syntactic considerations, but also on a host of other factors, both semantic (whether there is more than one semantically plausible interpretation) and pragmatic (the context and the expectations of the perceiver). Speakers and writers may take all of these factors into consideration, and adjust their syntax so as to avoid ambiguity whenever the danger arises. I will call this possibility 'tactical ambiguity avoidance'.

What I call 'strategic' and 'tactical' ambiguity avoidance are similar in that both involve syntactic choice. The difference is that in the 'strategic' case, the factors involved are primarily syntactic (or lexical), whereas in the 'tactical' case they may include semantic and pragmatic considerations as well.

My focus in this paper is on strategic ambiguity avoidance—cases where producers have a choice between alternative syntactic constructions, and the role that syntactic and lexical factors might play in such choices. In what follows, I discuss two important sites of syntactic choice in English: the use of *that* in embedded clauses, and the use of complementizers or relative pronouns in object relative clauses. In the first case, I review research by others (on both speech and written language) who cast doubt on the ambiguity-avoidance hypothesis; I question whether the data really justify this conclusion. In the second case, I examine new data suggesting that ambiguity avoidance does, indeed, play an important role in syntactic choices, though in a rather different way than has previously been considered.

3. A CASE STUDY: *that* IN EMBEDDED CLAUSES. In English embedded clauses (also known as object clauses or sentential complements), the inclusion of the overt complementizer *that* is (for most verbs) optional:

(4) a. He said he was coming.
b. He said that he was coming.
c. She knew the story was true.
d. She knew that the story was true.

Whether or not to include *that* is thus a choice that English speakers constantly confront. It seems possible that one factor here would be the avoidance of ambiguity. Note that some verbs allowing embedded clauses, such as *know*, are also transitive, allowing a direct object. In that case, the omission of *that* could result in a 'garden-path' effect, in which the intended subject of the embedded clause was understood as a direct object (*she knew the story*). Possibly, the use of *that* is tailored to avoid such ambiguities. Clearly what is at issue here is not a general syntactic rule (always use *that* with embedded clauses), or even lexical rules (always use *that* after certain verbs), since the use of *that* is (with many verbs) optional. However, there might be a syntactic strategy involved, favoring the inclusion of *that* under certain syntactic circumstances that are otherwise likely to be ambiguous; or one might find tactical use of *that* simply to avoid ambiguity in specific cases.

Two studies have examined the use of *that* in embedded-clause (EC) constructions. Elsness (1984) studied embedded clauses in the Brown corpus, a large corpus of written text, and presents several conclusions about the factors involved in the use of *that*. Elsness notes, first of all, that in cases where an adverbial phrase occurs between the matrix verb and the embedded clause, *that* is much more likely to occur than in other cases—that is, 5a would be more likely than 5b:

(5) a. John said on Friday that he was coming.
b. John said on Friday he was coming.

Elsness notes the possibility that this may be intended to reduce ambiguity—specifically, an ambiguity as to whether the adverbial phrase belongs to the main clause or the embedded clause (such ambiguities could evidently arise in cases like 5b).

Another strong pattern emerging from Elsness's data is that *that* is less likely to be used when the subject of the embedded clause is a pronoun. (*That* is used in 25% of cases with pronoun subjects, 60% with other subjects.) Elsness observes that this, too, could be explained in ambiguity-avoidance terms: some pronouns have distinct nominative and accusative forms; thus there should be no danger of mistaking the embedded-clause subject for a direct object.

(6) a. He knows I . . . (am coming).
b. He knows me.

However, Elsness argues, the hypothesis that the relatively infrequent use of *that* with pronoun subjects is due to ambiguity avoidance suggests that it WOULD be used with pronouns like *you* that do not distinguish nominative and accusative forms. Yet no difference was detected in the use of *that* between cases where the embedded subject was a case-distinguishing pronoun and cases where it was not. Thus the ambiguity-avoidance explanation seems to fail in this case.

A related finding in Elsness's study is that *that* is less likely to be used in cases where the subject of the embedded clause refers to something already mentioned in the discourse. This might account for the fact that *that* is used less often with pronoun EC subjects, since pronouns are most likely to be anaphoric. Even among embedded clauses with pronoun subjects, Elsness observes, *that* is less likely to occur if the EC subject and the subject of the main clause are coreferential.[4] With other types of subject noun phrase, *that* is used more often with indefinite noun phrases as opposed to definite ones; and, Elsness argues, definite noun phrases are much more likely to be anaphoric (though they are not always). In terms of discourse structure, Elsness suggests that the omission of *that* may serve to bind the embedded clause more closely to the previous context—as would be appropriate if the clause contains anaphoric references—whereas the use of *that* detaches it.

Further support for the role of anaphoricity in the use of *that* with embedded clauses is provided in an experimental study by Ferreira and Dell (2000). Ferreira and Dell showed several sentences to subjects and then asked them to recall one of the sentences, given certain words as recall cues. In one experiment, the sentences given included one of the following (sometimes with *that* and sometimes not):

(7) a. I knew (that) I had booked a flight for tomorrow.
b. I knew (that) you had booked a flight for tomorrow.
c. You knew (that) I had booked a flight for tomorrow.
d. You knew (that) you had booked a flight for tomorrow.

The dependent variable of interest was whether subjects would include *that* in recalling the sentence. The authors followed the reasoning discussed earlier by Elsness: an ambiguity-avoidance strategy should result in greater use of *that* in cases where the following subject is the ambiguous *you*, rather than the unambiguous *I*. (Actually, as the authors point out, the danger of ambiguity only really arises in 7b; in 7d, *you* could not be mistaken for the direct object of the main clause, as the reflexive form *yourself* would be required in this case.) Ferreira and Dell also tested an alternative explanation for

[4] The difference between coreferential and noncoreferential pronouns was significant in one of two corpora examined by Elsness.

the use of *that* in EC situations: an 'availability' hypothesis. According to this hypothesis, the use of optional *that* is designed to give the speaker additional time to access the needed lexical item for the following subject; it is, then, a kind of placeholder designed to maintain speaking fluency. If the item at issue has already been used in the conversation, it is likely to be easily accessible; thus no placeholder is needed. This would predict a greater use of *that* when the EC subject is not the same as the matrix subject. The data supported the availability hypothesis. *That* was used significantly less often in coreferential sentences, such as 7a,d, than in noncoreferential ones. However, there was no specific preference for *that* in potentially ambiguous cases like 7b, again casting doubt on the ambiguity-avoidance hypothesis.

For Ferreira and Dell, then, the primary factor in the use of *that* is that it is preferred in cases where the EC subject is nonanaphoric; Elsness, too, seems to emphasize this as the most important factor. We should note, however, that these authors advance rather different explanations for this phenomenon. In Elsness's view, the reason for the preference for *that* with nonanaphoric EC subjects is informational. Using *that* conveys a certain separation between the EC subject and the previous discourse, perhaps cueing the perceiver that the subject is indeed not anaphoric; omitting it creates a 'closer clause juncture' and thus carries the opposite implication. Ferreira and Dell, by contrast, attribute the greater use of *that* with nonanaphoric subjects to constraints on on-line speech production: previously mentioned items are more readily available to the speaker and can be produced more quickly; thus no *that* is necessary. While this issue is not our main concern, it should be pointed out that the mere fact that the phenomenon at issue has been demonstrated in both speech and writing would seem to favor Elsness's explanation over Ferreira and Dell's. Ferreira and Dell's account centers on the speaker's need for fluency—the reluctance to disrupt the flow of words for even a split second. But it is hard to see why this argument would apply to WRITING. To the extent that the role of anaphoricity in the use of *that* proves real in both speech and writing, then, an information-based explanation seems more plausible than a production-based one.

Taken together, these studies would seem to warrant some skepticism as to the importance of ambiguity avoidance in syntactic choices. However, some important questions remain. In terms of the framework presented in §2 of this paper, these studies seem to focus mainly on LEXICAL syntactic strategies. Both Elsness (using written text) and Ferreira and Dell (using speech) found that there was no greater tendency to use *that* preceding the case-ambiguous pronoun *you*, as opposed to unambiguous pronouns like *I* or *she*. However, Elsness found a strong GENERAL tendency to omit *that* in cases where pronouns are used. It could well be argued that this, in itself, is an ambiguity-avoidance mechanism. Since most personal pronouns—all except *you* and *it*—are case-specific, whereas other noun phrases are not, it can be seen that embedded clauses with pronoun subjects (8a) will generally be less ambiguous than those without (8b).

(8) a. I believe I/she/we/they . . .
b. I believe the doctor/your report/Mary . . .

In short, the relative preference for *that* with nonpronoun subjects could be accounted for by a general strategy to this effect.

The strategy just proposed—favoring *that* with nonpronoun subjects—cannot account for all of Elsness's and Ferreira and Dell's findings. Both studies found that *that* is used less with pronouns that are coreferential with the main-clause subject than with other pronouns; Elsness also suggested that, among nonpronoun EC subjects, *that* was used less when the subject was anaphoric (though the data for this was inconclusive).

On the other hand, one important phenomenon observed by Elsness—the tendency to include *that* in embedded clauses with adverbial modifiers—is explained by the ambiguity-avoidance model and not the anaphoricity model. On balance, then, the evidence is inconclusive. In any case, there are grounds for further exploration of syntactic choices and how they are affected by ambiguity avoidance and other factors.

4. RELATIVE PRONOUNS OR COMPLEMENTIZERS WITH OBJECT RELATIVE CLAUSES. To gain further insight into factors guiding syntactic strategies, it seemed worthwhile to consider another common situation of syntactic choice. In English relative clauses, either the complementizer *that* or a relative pronoun such as *who* or *which* is obligatory in cases where the antecedent or 'matrix' noun phrase is the implied subject of the relative clause; see 9a,b. However, when the matrix noun phrase is the implied object of the relative clause, the relative pronoun or complementizer is optional (9c,d).[5]

(9) a. The dog that/who/which chased me was black.
b. *The dog chased me was black.
c. The dog that/who/whom/which I chased was black.
d. The dog I chased was black.

As noted earlier, it has been suggested that this rule in itself might reflect pressures of ambiguity avoidance; 9b, if permitted, might easily mislead the perceiver into thinking that *chased* was the main verb. In light of this, it is natural to wonder whether the choice of including or omitting a complementizer or relative pronoun in object relative clauses is guided in part by ambiguity avoidance.[6]

Two points must be clarified. First, our concern here is only with restrictive relative clauses. In nonrestrictive object relative clauses—*The dog, which I chased, was black*—a relative pronoun is obligatory, so there is no choice to be made. Second, I do not distinguish between the use of relative pronouns and complementizers; the models to be considered appear to make the same predictions for either one. The choice at issue, then, is between either (i) an overt relative pronoun or complementizer (I abbreviate this as 'RP/comp'), or (ii) nothing (sometimes indicated as 'Ø').

As with embedded clauses, the potential for ambiguity with object relative clauses would appear to depend on the nature of the relative clause (RC) subject, but in a rather different way. Consider the abstract syntactic structures shown in Table 1—possible object RC constructions with no RP/comp—and the examples of each one shown at right. Let us assume that, in the normal case, the antecedent NP involves at least a common noun (restrictive relative clauses attached to pronouns or proper nouns are extremely rare); it may or may not involve a determiner. Suppose the relative clause

[5] In no case is it permitted to include both a complementizer and a relative pronoun: **The dog that which chased me was black*. Some earlier stages of English did allow 'which that' (sometimes known as a 'doubly-filled complementizer'), as do some dialects of English today; our concern here is only with standard English.

[6] Little research has been done relating specifically to relative pronoun use in object relative clauses. Several statistical studies have been done regarding relative pronoun choice across all types of relative construction in English—subject, object, indirect object, prepositional, genitive, locative, and manner. These studies are largely concerned with patterns of usage across categories (Biesenbach-Lucas 1987, Kikai et al. 1987, Guy & Bayley 1995). Guy and Bayley note that null and *that* choices in restrictive relative clauses are more common in speech, while WH-forms are relatively favored in writing. Another factor cited by both Kikai and colleagues and Guy and Bayley is the animacy of the antecedent: WH-forms are biased towards animate subjects, Ø less so, and *that* least of all. Guy and Bayley also observe that WH- and *that* are favored over Ø in cases where the relative clause is separated from the antecedent, echoing Elsness's finding with regard to embedded clauses; the authors point out that this may reflect considerations of parsing. (The latter two points are considered further in §5.)

CATEGORY	MATRIX NP	RELATIVE CLAUSE	EXAMPLES
A	(D) N	Pronoun	1. The lawyer I . . .
			2. The lawyer you . . .
			3. We saw the lawyer I . . .
			4. Lawyers I . . .
B	(D) N	D N	5. The lawyer the company . . .
			6. This is the lawyer the company . . .
C	(D) N	N	7. The lawyer companies . . .
			8. The car companies . . .
			9. A lawyer companies . . .
			10. Lawyers companies . . .
			11. The lawyer big companies . . .
D	(D) N	Proper NP	12. The lawyer Smith . . .
			13. The lawyers Smith . . .
			14. A lawyer Smith . . .

TABLE 1. Possible relative clause constructions.

begins with a pronoun (category A in Table 1). In general, the syntactic structure will now be clear; it more or less has to be an object RC situation. This holds true whether or not the matrix NP is in subject position (as in items 1 and 2 in Table 1) or object position (as in item 3). Notice also that in this case (unlike the EC case), it makes no difference whether the pronoun is case-specific (as in item 1) or not (as in item 2); ambiguity does not arise in either situation. Nor is the situation affected by whether the matrix noun phrase has a determiner (compare items 1 and 4).

Consider also the case where the RC subject involves a determiner followed by a noun (category B). The situation is the same as with a pronoun subject: there seems little danger of ambiguity, whether or not the matrix subject carries a determiner, and whether or not the matrix NP is in subject or object position. The same holds true if the RC subject noun is preceded by a possessive pronoun (*The lawyer my company . . .*) or quantifier (*The lawyer many/both/some companies . . .*). For convenience, these kinds of words are all classed as determiners in what follows. (When used in a stand-alone fashion, quantifiers like *many* and *some* can be considered pronouns, and are thus grouped in category A.)

Now consider the case where the RC subject is a plural or mass noun, without a determiner (in the case of singular count nouns, a determiner is of course obligatory). In this case, the risk of ambiguity is very real: the matrix NP and RC subject might seem to form a single noun phrase, with the head noun of the matrix NP being a noun modifier to the RC subject (items 7 and 8). This could occur regardless of the position of the matrix noun phrase in the sentence. It does, however, depend on a number of other syntactic factors. (Of course, semantic factors could also affect the situation—for one thing, some combinations of noun modifier and noun are more plausible than others; I return to this issue in a later section.) If the matrix noun phrase has a singular determiner, as in item 9, no ambiguity occurs, as this determiner is incompatible with the plural/mass RC subject. If the matrix noun phrase is plural (item 10), ambiguity becomes less likely, as noun-modifiers are rarely plural; however, plural noun-modifiers are far from unheard of (consider cases like *weapons violations*, *human rights issues*, *narcotics officer*, *appeals court*, *securities firm*, and *futures market*). An adjective preceding the RC subject (item 11) would also reduce ambiguity, though again not eliminate it; noun phrases in which a noun-modifier precedes an adjective do sometimes occur (consider phrases like *city clerical worker* and *object relative clause*).

A final general case involves relative clauses with proper-noun subjects (items 12–14). Here, too, ambiguity may arise: the matrix NP and RC subject may join together into a single noun phrase. Such constructions—sometimes called restrictive appositives (Quirk et al. 1985)—are fairly frequent in phrases like *the actor Paul Newman* and *the accounting firm Smith and Jones*. Again, various factors could eliminate the ambiguity; if the matrix noun is plural (item 13), or preceded by an indefinite determiner (item 14), the appositive interpretation becomes much less plausible.[7]

To summarize, the risk of ambiguity is greatly affected by the nature of the relative clause subject. Any relative clause can be categorized into one of the four categories just presented, according to the structure of the RC subject: (A) pronoun; (B) common noun with determiner; (C) common noun with no determiner; (D) proper noun or noun phrase. The ambiguity-avoidance hypothesis predicts that RP/comp would be included more often in categories C and D, less often in categories A and B.

Informal inspection of written text from various sources suggested that this hypothesis might be borne out. The following examples (all taken from the *New York Times*) are illustrative; each one contains an object relative clause in which RP/comp is included. (Italics have been added to indicate the segment of interest.)

(10) a. *The biological toll that logging* can take on a landscape is well known, a toll that is especially harsh on a forest hit hard by high-intensity fires.
b. The Senate bill would require companies either to forgo *the tax deduction that corporations* take when employees exercise options or to include the cost of options, as estimated by a complex formula, in their income statement.
c. The company's sensor chip is being used in *a single-lens reflex camera that Sigma, a Japanese camera and lens maker*, plans to begin selling for about $3,000 later this month.
d. It is a harsh comeuppance for *the company that Bernard J. Ebbers* first outlined on a napkin in 1985, expanded through dozens of acquisitions and built into one of the most prominent success stories of the 1990's.

(Examples 10a,b would belong to category C; 10c,d would belong to category D.) Consider, in each case, how the sentence would read if *that* were omitted. The intended meaning of the sentence could probably still be inferred, through the use of semantic and pragmatic information. But the omission of *that* produces a cluster of lexical items—*biological toll logging, single-lens reflex camera Sigma*—which could certainly cause at least momentary confusion or misinterpretation. It seems plausible that the function of RP/comp in object relative clauses, at least in part, is to reduce the potential for such confusions.[8]

The categorization system proposed here might be questioned. As mentioned earlier, ambiguity is also affected by more fine-grained syntactic distinctions within these categories: for example, in category C, whether the RC subject noun is preceded by an

[7] A commonly arising complication here is actual titles, such as *Mr.* If such a title is used before the RC subject, it makes it clear that the phrase is anaphoric, and thus an appositive interpretation is unlikely: one would not say *The actor Mr. Newman attended the event.* However, proper NPs with titles were included in category D nonetheless.

[8] Actually, in some cases, the ambiguity is not just of the 'garden-path' kind, but persists until the end of the sentence. In 10b, for example, if *that* is omitted, the sentence remains ambiguous (at least syntactically) even when seen in its entirety: it could be that *the tax* is the matrix NP and *deduction corporations* is the RC subject.

adjective, or whether the matrix NP carries a singular determiner. But this is not a fatal problem for the proposed model. Even if not all determinerless common-noun RC subjects result in ambiguities, it may well be the case that they sometimes cause ambiguities (more often than, say, pronouns), and that a general strategy of using RP/comp with such subjects would be, on balance, advantageous. In any case, this categorization system provides a starting point. If the predictions offered here are supported, this will present a phenomenon to be explained; we may then consider whether it is best explained by the current ambiguity-avoidance model, some other ambiguity-avoidance model, or some other model entirely.[9]

In examining the use of RP/comp in object relative clauses, I consider not only the ambiguity-avoidance model, but also the 'anaphoricity' hypothesis discussed earlier. While neither Elsness nor Ferreira and Dell discuss RP/comp usage in relative clauses, the prediction of the anaphoricity hypothesis seems clear: by this model, the use of RP/comp should be determined by whether the subject of the relative clause is anaphoric with items presented earlier in the discourse. By either Elsness's information-based motivation for the anaphoricity hypothesis or Ferreira and Dell's 'availability' motivation, it is difficult to see why this hypothesis would apply to embedded clause subjects but not to relative clause subjects.[10] (As noted earlier, the availability explanation seems less plausible in general with regard to written language.)

In short, the question is this: is the use of RP/comp in object relative clauses guided more by ambiguity avoidance—reflected in the syntactic structure of the RC subject phrase—or by anaphoricity? The situation is complicated by the fact that the anaphoric status of a noun phrase and its syntactic structure are far from independent. Consider the categories in Table 1. Pronouns (category A) are anaphoric in the vast majority of cases; third-person pronouns are always anaphoric. Nouns with determiners (category B) are also most likely to be anaphoric, especially if the determiner is a definite one (as it most often is). Plural noun phrases without determiners (category C) might or might not be anaphoric; something like item 7 in Table 1 could conceivably occur whether or not *companies* had been referred to previously in the discourse. Proper noun phrases also (category D) may or may not be anaphoric. Thus the predictions of the anaphoricity model and the ambiguity-avoidance model may to a large extent coincide. For this reason it is important to examine the independent contributions of the two models.

5. RP/COMP IN OBJECT RELATIVE CLAUSES: A CORPUS STUDY. The use of RP/comp in object relative clauses was examined statistically, using written text. The corpus used was the Penn Treebank, a corpus of over one million words from the 1989 *Wall Street*

[9] Under this system, noun phrases are categorized according to their head word (along with their determiner, if any). It might be argued that they should be categorized in terms of their first word instead: if the first word is a determiner, quantifier, or adjective, the phrase is not ambiguous; if the first word is a common or proper noun, it is ambiguous. Actually, these two systems would yield very similar results. In NPs involving a determiner, the determiner virtually always is the first word; in proper noun phrases, the first word most often is a proper noun. However, a 'first-word'-based system would yield different results in some cases and might be worth considering. The question is, of course, which system is most plausible as a model of the strategies of language-users; this is difficult to know.

[10] It is true that relative clauses may involve either a complementizer or a relative pronoun, whereas embedded clauses may only use a complementizer; perhaps this would make a difference for some reason. In practice, however, object relative clauses appear to involve complementizers rather than relative pronouns in the vast majority of cases (at least in newspaper text), as shown in §5.

Journal, annotated with syntactic information.[11] The first 6 out of 25 segments of the corpus (12,948 sentences) were analyzed, and all object relative clauses were identified, using a computer program especially designed for the purpose. An example of a sentence from the corpus containing an object relative clause is shown in 11:

(11) ((S (PP-TMP For (NP the fifth consecutive month)), (NP-SBJ purchasing managers) (VP said (SBAR 0 (S (NP-SBJ (NP prices) (PP for (NP (NP the goods) (SBAR (WHNP-1 0) (S (NP-SBJ they) (VP purchased (NP *T*-1)))))))) (VP fell)))) .))

The crucial information to notice is that there is an NP, beginning with *the goods*, containing another NP followed by an SBAR (i.e. an S-bar, S′). The SBAR begins with a WHNP (relative pronoun or complementizer); in the present case this contains only '0' (a null element), but if an RP/comp was included it would occur in that position. Following this is an S, containing an 'NP-SBJ'—a subject noun phrase—that contains a word. If the relative clause were of the subject type, the NP-SBJ would not include a word but rather '*T', indicating another null (trace) element. In short, an object relative clause can be identified as an NP containing first an NP and then an SBAR, which in turn contains a WHNP followed by an S whose NP-SBJ contains a non-null element. That this algorithm reliably identifies all object relative clauses was confirmed by consultation of the documentation for the treebank, as well as by comparison with the results of a manual analysis of one section of the treebank. (One exception is conjoined relative clauses—*The dog that I chased and that I liked*; these are not identified by the algorithm, and were thus not included in the present study. The algorithm also identifies a small number of nonobject relatives, such as possessive relatives; these were weeded out by hand after running the program.)

One problematic case involves relative clauses that themselves include embedded clauses; these can be either subject type (12a,b) or object type (12c,d). RP/comp is optional in either situation:

(12) a. The dog you said chased you is here.
b. The dog that you said chased you is here.
c. The dog you said you chased is here.
d. The dog that you said you chased is here.

Under both the anaphoricity model and the ambiguity-avoidance model, it appears that the same predictions would be made for embedded relatives (whether subject or object) as for ordinary object relatives. For that reason, all such relative clauses were included in the data set. Prepositional relatives (*the man I talked to*) were also included (but not ones with fronted prepositions, e.g. *the man to whom I talked*: in this case, the relative pronoun is obligatory). Given these rules, categorization was unproblematic; it was always clear whether a given case should or should not be included in the data set.

This analysis produced a corpus of 391 relative clauses. These were inspected individually, and it was found that 62 of them were in quoted statements. These were excluded

[11] See Marcus et al. 1993; further documentation is provided at http://www.cis.upenn.edu/~treebank. The corpus also includes some text from other sources, but only the *Wall Street Journal* portion of the treebank was used in the current study.

One might wonder whether the use of RP/comp in the *Wall Street Journal* is governed by an explicit editorial policy. In response to a query, the style guide editor of the *Journal* indicated only that *that* is the preferred form for restrictive relative clauses, while Ø is also permissible; there appears to be no more specific policy beyond this.

for two reasons. First, it seemed wiser to include written text only, as spoken and written language may differ in certain respects; secondly, in a quotation sentence (which may be taken out of a longer discourse), it is often impossible to tell if a given element is anaphoric.[12] The remaining 329 relative clauses constituted the data set for the study. Of these, 90 involved a relative pronoun or complementizer (27.4% of the total); 239 did not. For those that did, *that* was used in 87 cases; *whom* was used in 3 cases; no other relative pronouns were used. The clauses were analyzed by hand in two respects. First, they were categorized as to the structure of the RC subject phrase according to the system presented in Table 1.[13] Next, the clauses were analyzed as to whether the subject phrase was anaphoric. Recall that the treebank consists of excerpts from newspaper items (the beginning and ending of excerpts is marked in the treebank); while not all excerpts include the entire news item, all excerpts contain a continuous portion of the item starting at the beginning.[14] Thus it could usually be determined whether or not something had been previously referred to in the item. There were sometimes problems, however, in determining whether two discourse elements were coreferential. About 2% of cases were considered questionable; these were simply categorized into one category or another as consistently as possible.[15]

One might expect that in object relative clauses, subject phrases would most often be anaphoric. Generally, restrictive relative clauses are used to relate a new discourse element (the matrix NP) to elements already present in the discourse; in such cases the RC subject should generally be anaphoric. This proved to be the case; out of 329 cases, the subject NP was anaphoric in 277 (84.2%). All third-person pronouns were simply assumed to be anaphoric; other pronouns were checked individually, however, and a small number of first- and second-person pronouns proved to be nonanaphoric.

	A (Pronoun)		B (D+N)		C (No D+N)		D (ProperN)	
RP/comp	17	(10.9%)	18	(25.0%)	32	(84.2%)	23	(36.5%)
No RP/comp	139	(89.1%)	54	(75.0%)	6	(15.8%)	40	(63.5%)
TOTAL	156	(100.0%)	72	(100.0%)	38	(100.0%)	63	(100.0%)

TABLE 2. Use of RP/comp in object relative clauses (by RC subject category).

Table 2 shows the aggregate data regarding the use of RP/comp and the syntactic-structure category of the RC subject phrase. A strong relationship is seen, in exactly the predicted direction. In proportional terms, RP/comp is included least often with pronouns and determiner + noun constructions; it is included more often with proper-noun phrases, and most often with determinerless common-noun phrases. The relationship between these variables is highly significant ($\chi^2 = 85.93$, $p < 0.0001$). Table 3 shows the aggregate data regarding the use of RP/comp and anaphoricity. Here, too, a highly significant relationship is found: RP/comp is used much more often in non-anaphoric cases ($\chi^2 = 59.62$, $p < 0.0001$).

[12] In cases where part of a relative clause was in quotes, the clause was included if it appeared that the decision whether to use an RP/comp rested with the writer rather than the speaker.

[13] This yielded only two questionable cases. One was the noun phrase *only a very few*; the other was *we Southerners*. Both were classed in category A.

[14] This rule is not stated in the documentation for the treebank, but inspection of a large number of examples confirmed it in every case, and it was therefore assumed to be valid in general.

[15] For example: In the phrase *the minimum amount of cash an investor must put up* (where *investors* has been used before), is *an investor* anaphoric? Or another example: when are the pronouns *one* and *nobody* anaphoric? See Hirst 1981 and Carter 1987 for discussion of some issues that arise in determining anaphoricity.

	ANAPHORIC	NONANAPHORIC
RP/comp	53 (19.1%)	37 (71.1%)
No RP/comp	224 (80.9%)	15 (28.9%)
TOTAL	277 (100.0%)	52 (100.0%)

TABLE 3. Use of RP/comp in object relative clauses (by anaphoricity).

Next, the contributions of anaphoricity and RC subject category were examined independently. The relationship between RC subject type and use of RP/comp was analyzed for anaphoric and nonanaphoric cases separately (Table 4). For anaphoric

	A (Pronoun)	B (D + N)	C (No D + N)	D (ProperN)
ANAPHORIC				
RP/comp	16 (10.5%)	6 (11.5%)	15 (78.9%)	16 (29.6%)
No RP/comp	136 (89.5%)	46 (88.5%)	4 (21.1%)	38 (70.4%)
NONANAPHORIC				
RP/comp	1 (25.0%)	12 (60.0%)	17 (89.5%)	7 (77.8%)
No RP/comp	3 (75.0%)	8 (40.0%)	2 (10.5%)	2 (22.2%)

TABLE 4. Use of RP/comp by RC subject category (for anaphoric/non-anaphoric cases).

cases, a highly significant relationship was found ($\chi^2 = 56.99$, $p < 0.0001$); for non-anaphoric cases, the relationship was also significant, though only narrowly so ($\chi^2 = 8.66$, $p < 0.05$) (note the relatively small body of data here). In both cases, the pattern is as expected: RP/comp is used less often with pronouns and determiner + noun phrases, more often with proper nouns and determinerless noun phrases (see Figure 1).

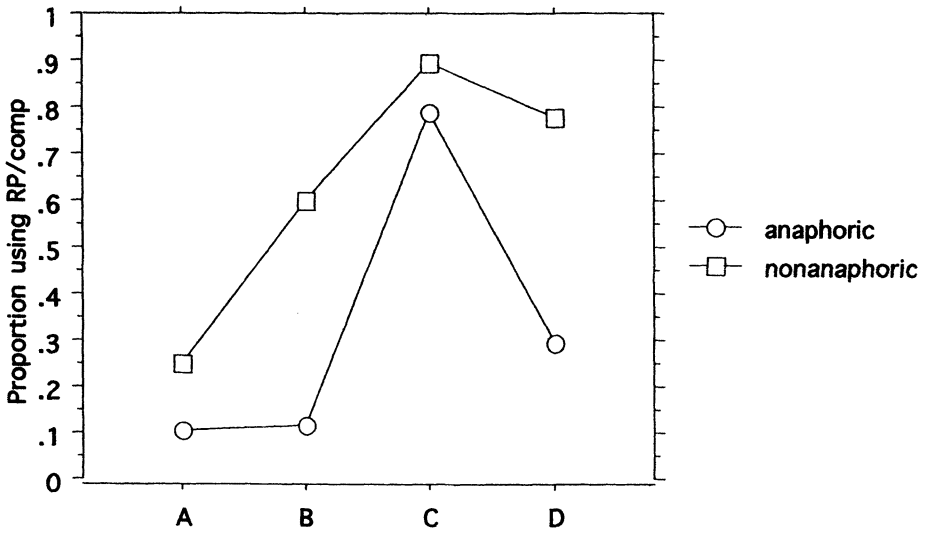

FIGURE 1. Use of RP/comp by RC subject type.

Examining the data for the categories predicted to be ambiguous (C and D) and unambiguous (A and B), we can see that the difference between category C and the unambiguous categories is greater than that between category D and the unambiguous categories. (This was revealed statistically as well. When category C alone was compared to the union of categories A and B, a significant difference was observed for

both anaphoric and nonanaphoric cases; when category D was compared to the union of categories A and B, the difference was significant for anaphoric cases only.)[16] This difference, too, might well be attributed to ambiguity avoidance. The competing syntactic structure that (according to the current model) might be inadvertently implied by a proper-noun RC subject (category D) is the restrictive appositive construction: something like *the accounting firm Smith & Jones*. This construction is relatively rare, and the danger of ambiguity it presents may not be as severe as in the case of determinerless common-noun phrases (category C), where the competing construction—a plural or mass noun preceded by a noun-modifier—is much more common.

Table 5 shows the relationship between the use of RP/comp and anaphoricity, within each syntactic-structure category (see also Figure 2). In two of the four categories, a significant relationship was observed: determiner + noun ($\chi^2 = 18.09$, $p < 0.0001$) and proper nouns ($\chi^2 = 7.71$, $p < 0.01$). For the other two categories, the relationship was in the expected direction but not significant (for pronouns, $\chi^2 = .84$, $p = 0.36$; for determinerless NPs, $\chi^2 = .79$, $p = 0.37$). Again, the relatively small data sets for these categories should be borne in mind.

	ANAPHORIC		NONANAPHORIC	
A (Pronoun)				
RP/comp	16	(10.5%)	1	(25.0%)
No RP/comp	136	(89.5%)	3	(75.0%)
B (D + N)				
RP/comp	6	(11.5%)	12	(60.0%)
No RP/comp	46	(88.5%)	8	(40.0%)
C (No D + N)				
RP/comp	15	(78.9%)	17	(89.5%)
No RP/comp	4	(21.1%)	2	(10.5%)
D (ProperN)				
RP/comp	16	(29.6%)	7	(77.8%)
No RP/comp	38	(70.4%)	2	(22.2%)

TABLE 5. Use of RP/comp by anaphoricity, for different RC subject categories.

The general conclusion seems clear: both ambiguity avoidance and anaphoricity contribute independently to the use of RP/comp in object relative clauses. Even when only nonanaphoric cases were considered, RP/comp was significantly more likely to be used with some syntactic categories of RC subject than others; when anaphoric cases alone were considered, the same relationship emerged even more strongly. And in both cases, the pattern was just as predicted by the ambiguity-avoidance model. On the other hand, when only sentences within a certain syntactic-structure category were considered, a significant bias towards use of RP/comp in nonanaphoric cases emerged in two out of four categories.

In order to assess the relative importance of anaphoricity and RC subject on the use of RP/comp, a logistic regression test was performed. For the purposes of this analysis, the four categories of RC subject were collapsed into two, A plus B and C plus D. All three variables (use of RP/comp, anaphoricity, and RC subject) were coded as dummy variables with values of either 0 or 1. The test yielded coefficients of -2.280 for

[16] For the comparison between category C and the union of categories A and B: for anaphoric cases, $\chi^2 = 58.35$, $p < 0.0001$; for non-anaphoric cases, $\chi^2 = 6.27$, $p < 0.05$. For the comparison between category D and the union of categories A and B: for anaphoric cases, $\chi^2 = 12.07$, $p < 0.001$; for non-anaphoric cases, $\chi^2 = 1.53$, $p = 0.22$.

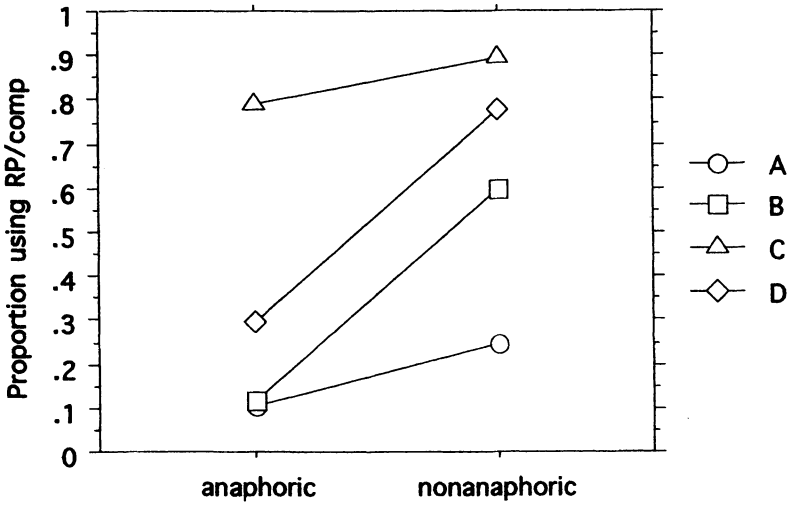

FIGURE 2. Use of RP/comp by anaphoricity.

anaphoricity ($p = 0.000$) and 1.625 for relative-clause subject ($p = 0.017$); the interaction between the two variables was not significant. This confirms the earlier conclusion that both anaphoricity and relative-clause subject are factors in RP/comp usage, and suggests that anaphoricity carries slightly greater weight.

The data regarding the use of RP/comp in object relative clauses was analyzed with respect to three other factors. First, the relationship between the use of RP/comp and the length of the relative clause was examined. It seemed possible that RP/comp would be used more often before longer relative clauses—perhaps as a perceptual cue that a long dependent clause is coming. If so, the findings reported earlier might be an artifact of this tendency; for example, pronoun RC subjects might be associated with shorter relative clauses. However, analysis of the data suggested that this concern was unfounded. The average length of relative clauses with RP/comp was 8.38 words; the average length of those without RP/comp was 7.13 words, a difference just short of statistical significance ($F(1,327) = 3.35, p = 0.07$).[17] A second factor considered was animacy of the antecedent. Earlier data on relative pronoun use (Kikai et al. 1987, Guy & Bayley 1995) suggests that in general, WH-forms are more favored with animate antecedents, while *that* is more favored with inanimate antecedents, with Ø in between. While WH-forms are almost nonexistent in the current data, it seemed possible that animacy would be a factor in the choice between *that* and Ø. Again, however, analysis of the data suggested otherwise; the relationship between animacy of the antecedent and RP/comp choice did not approach significance ($\chi^2 = .24, p = 0.63$).

A third and more significant factor concerned cases where the head word of the matrix NP and the relative clause are separated. The most common cause of this is a

[17] Regardless of the level of significance, the small difference in length found here (1.25 words) seems to argue against RC length per se as an important factor in the use of RP/comp. Even if RC length were not directly a factor, we would expect some difference in average length between RP/comp and Ø cases. As noted earlier, pronoun subjects are strongly associated with no RP/comp, and since pronoun subject phrases are almost always just one word long (unlike other subject phrase types), relative clauses containing them no doubt tend to be somewhat shorter than others on average. Quite possibly the small difference in length between RP/comp and Ø cases is mainly due to this factor.

prepositional phrase modifying the matrix NP, as in these two sentences from the treebank:

(13) a. Alltel Corp. said it will acquire the 55% of Pond Branch Telephone Company Inc.'s cellular franchise that it doesn't own already.
b. Commonwealth Edison now faces an additional court-ordered refund on its summer/winter rate differential collections that the Illinois Appellate Court has estimated at $140 million.

Analysis of the object RC data set showed that 46.9% of cases with 'interruptions' (as I call them) included an RP/comp, as opposed to only 25.2% of those without; this difference proved to be statistically significant ($\chi^2 = 6.80, p < 0.01$). The possibility that this factor was confounded with either RC subject type or anaphoricity was ruled out by tests showing that neither of these variables is significantly correlated with the presence of 'interruptions' between the matrix NP head word and the relative clause.

What could account for the increased use of RP/comp with interrupted relative clauses? The ambiguity-avoidance hypothesis is one possibility. Recall that Elsness (1984) pointed to ambiguity avoidance to explain a similar phenomenon, the increased use of *that* in embedded clauses in cases where the matrix verb and embedded clause are separated. In that case, the complementizer serves a clear disambiguating function; without it, it would be unclear whether the adverbial phrase applied to the main clause or the embedded one. The danger of ambiguity arises with interrupted RC constructions, too: in many cases, the relative clause could attach either to the intended matrix NP (*the 55%* in 13a) or to the PP object (*Pond Branch Telephone Company Inc.'s cellular franchise*). In this case, however, it is not so clear how the use of RP/comp removes the ambiguity; even with RP/comp, the relative clause could still attach to either NP. One possibility is that there is simply a convention, understood by producers and perceivers, that when RP/comp is used following an NP PP construction, the relative clause generally applies to the NP further back rather than to the more recent one.

6. GENERAL DISCUSSION. It is useful to bring together these findings about RP/comp in relative clauses with the earlier findings about *that* in embedded clauses. Regarding embedded clauses, it was noted that one important finding—the preference for *that* in cases with adverbial modifiers—seems most readily explained by the ambiguity-avoidance hypothesis. Another, the preference for *that* with nonpronoun as opposed to pronoun subjects, can be explained by either the anaphoricity model or the ambiguity-avoidance model, providing we allow for the possibility of a general (not word-specific) syntactic strategy distinguishing only between pronouns and nonpronouns. A third finding, the preference for *that* with nonanaphoric subjects even within pronoun and nonpronoun categories, seems most compatible with an anaphoricity explanation. The most natural conclusion from this would seem to be that both anaphoricity and ambiguity avoidance can influence syntactic choices. The data just presented regarding relative clauses point to exactly the same conclusion. Anaphoricity seems to be a factor in the use of RP/comp even within RC subjects of a certain syntactic type (at least, within some categories of RC subject). But even when anaphoricity is controlled, there emerges a preference for RP/comp with some kinds of subject types over others; and the ambiguity-avoidance account seems to provide the most plausible explanation for this.

In terms of the framework proposed at the beginning of the study, there seems to be evidence for ambiguity avoidance as a factor in syntactic strategies. However, the evidence presented so far seems to point entirely to general syntactic strategies rather than lexical ones: that is to say, ambiguity avoidance seems to be influenced more by

general considerations of syntactic structure than by lexical distinctions. Some evidence for this was presented earlier: while writers tend to use *that* less often with pronoun subjects, as shown by Elsness, neither Elsness nor Ferreira and Dell found any difference in this regard between seemingly ambiguity-prone pronouns like *you* and unambiguous ones like *I*. Ferreira and Dell found further evidence against ambiguity avoidance at the lexical level (not mentioned in my earlier discussion of their study). In one experiment, subjects had to recall reduced relative constructions, in which speakers have a choice whether or not to include the relative pronoun and finite copula: for example, *The astronauts (who were) selected/chosen made history*. Ferreira and Dell reasoned that if speakers were guided by ambiguity avoidance, they might be more likely to use *who were* with a verb like *selected* rather than one like *chosen*, since *selected* can also be a simple past form (*the astronauts selected their flight plan . . .*). However, no such effect was found; speakers did not tend to use a full relative construction more with ambiguous verbs.

Ferreira and Dell provide yet more evidence against ambiguity avoidance at the lexical level. The authors note that while some verbs taking embedded clauses are also transitive (like *announce*), others are not (like *insist*):

(14) a. I announced (that) the meeting would be on Tuesday.
b. I announced the meeting.
c. I insisted (that) the meeting would be on Tuesday.
d. *I insisted the meeting.

If ambiguity avoidance were indeed a factor at the lexical level, speakers might be more likely to use *that* with verbs like *announce* that might otherwise cause ambiguity, as opposed to verbs like *insist* where ambiguity seems unlikely. To test this, the authors found statistical data regarding the frequency of verbs taking different complements such as direct objects and embedded clauses, and assigned each verb a measure of its bias towards EC complements. This was then compared with the tendency of subjects in Ferreira and Dell's experiments (reported earlier) to use *that* following a given verb. No significant correlation was found, suggesting that the bias of certain verbs towards non-EC complements does not pressure speakers to use *that* more frequently. I examined the same question regarding the Penn Treebank data: in embedded-clause situations, is *that* used more frequently with verbs that are also used transitively much of the time? The data is shown in Figure 3 (the entire *Wall Street Journal* portion of the treebank was used for this experiment). Only verbs that frequently occur in EC situations (40 times or more in the treebank) are shown—a total of 26 verbs. The proportion of EC occurrences of the verb using *that* is plotted against the proportion of the time that the verb is used transitively. A bias towards *that* with often-transitive verbs would predict a positive relationship; a positive relationship is found, but only a very small (nonsignificant) one ($r = .136$, $p = 0.51$). While verbs like *say* and *think* fit the hypothesized pattern, verbs like *argue* and *agree* go against it; they are almost never used transitively, yet are almost always used with *that*.[18]

[18] The data was also analyzed in another way: for each verb, I examined the relationship between the tendency to include *that* in embedded clauses and the proportion of occurrences of the verb in which it takes an embedded clause (as opposed to other usages: intransitive, transitive, indirect question, etc.). The ambiguity-avoidance hypothesis might predict that *that* would be used more with verbs that were used less often in embedded clauses. Again, a relationship in the expected direction was found, but it was very small and not significant.

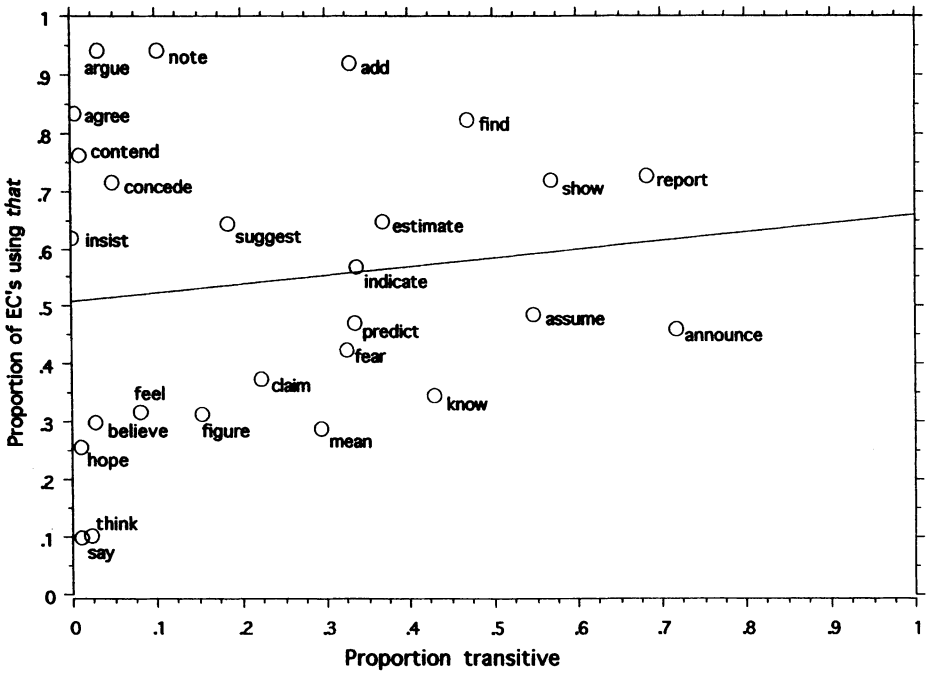

FIGURE 3. Verbs taking embedded clauses in the Penn Treebank.

In short, studies of several quite different situations (verbs in reduced relative clauses, verbs taking embedded clauses, and pronouns in embedded-clause subjects) have failed to produce any effect of ambiguity avoidance at the lexical level.[19] If ambiguity avoidance fails to emerge at this level, this seems to cast doubt as to whether it might occur at still more specific, situational levels, under the influence of semantic and pragmatic factors—what I referred to earlier as 'tactical ambiguity avoidance'. For example, suppose you wanted to say *I announced (that) X was* . . . , where *X* is a noun phrase. If tactical ambiguity avoidance were a factor, your decision to include *that* would depend on whether you thought your listener was in any danger of thinking that you intended to say *I announced X*, with *X* as a direct object. This might depend on whether *X* is a plausible direct object for *announce* (*the meeting* is, while *the corporation* is not); it might also depend on whether *I* is a plausible subject for the sentence in question, in general or in that specific situation. One reason to doubt that this kind of tactical ambiguity avoidance takes place is that, if it did, we would expect it to show up at the lexical level. Presumably such situation-specific ambiguities arise more often with *announce* than with *insist* (they never arise with *insist*, since *insist* cannot be transitive); if tactical ambiguity avoidance were at work, we would therefore expect that *that* would occur more often with often-transitive verbs, even if there was no general strategy to this effect. Thus the apparent absence of ambiguity avoidance at the lexical level strongly argues against its reality at more situation-specific levels as well.

Since there appears to be significant evidence for ambiguity avoidance as a factor in syntactic strategies, it is appropriate to take another look at the prima facie plausibility of this idea. First of all, it is hardly surprising that ambiguity avoidance is more promi-

[19] A recent study by Wasow and Arnold (2003) provides yet more support for this conclusion.

nent at the general-strategic level rather than at the lexical-strategic or tactical level. Avoiding ambiguity is hard: it requires thinking about how your sentence or a portion of your sentence might be misconstrued. The more kinds of information are taken into account in this process, the harder it will be. Considering lexical information (e.g. the fact that *selected* is more ambiguous than *chosen*, or the fact that *you* is more ambiguous than *I*) could be very computationally demanding; considering pragmatic factors, even more so. However, the use of general syntactic strategies (such as 'use *that* in relative clauses when the RC subject is determinerless') seems much more feasible—particularly if the number of strategies is relatively small, and the conditions for their application are defined in a fairly simple way. One might also ask whether it would even do any good to incorporate semantic or pragmatic factors in syntactic choices. If people do not use such information in PARSING language, there is no point in considering it when trying to avoid ambiguity. The extent to which nonsyntactic information is used in parsing is one of the perennial questions of psycholinguistics; while this is hardly a resolved issue, there is considerable evidence that the initial interpretation of a sentence is guided primarily by syntactic factors (Mitchell 1994:388–99). If so, it may be that bringing semantic or pragmatic information to bear in avoiding ambiguity would have little benefit, at least with regard to the initial stage of parsing.

One might wonder whether ambiguity avoidance operates differently between spoken and written language. In one respect, the need for ambiguity avoidance may actually be greater in speech than in writing, due to the greater time pressures on comprehension (as well as other hindrances—distractions, competing sound sources, and the like). In another respect, we might expect to find ambiguity avoidance more in written language, since writing allows more time for such considerations to be brought to bear (although once again, the kind of strategic ambiguity avoidance hypothesized here may not be especially demanding). In any case, it cannot be assumed that the phenomena observed here in writing would necessarily be reflected in speech.[20]

It was observed earlier that syntactic choices are a well-documented phenomenon in language. In some cases they appear to be guided by considerations other than ambiguity avoidance. For example, it has been shown that in the case of V NP PP constructions, speakers tend to place the NP after the PP in cases where it is 'heavier'—longer, essentially (Arnold et al. 2000).[21] Another factor in syntactic choices is what has been called 'syntactic persistence' or 'syntactic priming'. In choosing between 'double object' uses (15a) and 'prepositional object' uses (15b) of ditransitive verbs, speakers favor the usage they have recently heard or used.

(15) a. The girl gave the boy the book.
b. The girl gave the book to the boy.

This has been demonstrated both for a single speaker (where the syntax of sentences read aloud affects spontaneously generated sentences [Bock 1986]), and for dialogue situations (where one speaker's syntactic behavior influences another [Branigan et al. 2000]). The role of factors such as 'heaviness' and syntactic persistence in syntactic choices suggests that ambiguity avoidance is far from the only consideration guiding

[20] In a study of relative pronoun usage including both written and spoken data, Guy and Bayley (1995) note the tendency to include *that* with interrupted RCs (as has been observed here for written data), and suggest that this may reflect pressures of ambiguity avoidance. However, the authors do not analyze this phenomenon in spoken and written data separately.

[21] Speakers also tend to place the NP last if it involves new, rather than given, information, but 'heaviness' has been shown to play a role independently of 'newness'; see Arnold et al. 2000.

such choices. On the other hand, these phenomena also provide a kind of support for the plausibility of general syntactic strategies for ambiguity avoidance, by showing that similar strategies exist elsewhere—strategies in which one syntactic option or another may be chosen based on factors that, in themselves, are primarily syntactic.

At the most general level, the current study adds a small piece to the steadily (if slowly) growing body of evidence for ambiguity avoidance as an important determinant of syntactic structure and behavior. As noted earlier, ambiguity avoidance has received a number of scattered mentions in the syntactic literature, but has not received the focused and systematic attention it deserves. Relative clauses provide an instructive example of the explanatory power of ambiguity avoidance. As noted by Bever, the simple fact that relative pronouns or complementizers are required, in standard usage at least,[22] in subject but not object relative clauses is very naturally explained by an ambiguity-avoidance account. In addition, the current study suggests that ambiguity avoidance is an important factor in the use of RP/comp with object relative clauses. Ambiguity avoidance can account for the preference for RP/comp with some RC subject types over others, and also for the tendency to use RP/comp with interrupted relative clauses. Thus both the rules regarding the use of RP/comp and the variation within those rules seem to be attributable to ambiguity avoidance. Whether other phenomena would also be illuminated by an ambiguity-avoidance approach remains to be seen; at the very least, this approach warrants further study and attention.

REFERENCES

ARNOLD, JENNIFER E.; THOMAS WASOW; ANTHONY LOSONGCO; and RYAN GINSTROM. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. Language 76.28–55.

BEVER, THOMAS G. 1970. The cognitive basis for linguistic structures. Cognition and the development of language, ed. by John R. Hayes, 279–362. New York: Wiley.

BEVER, THOMAS G., and D. TERENCE LANGENDOEN. 1971. A dynamic model of evolution of language. Linguistic Inquiry 2.433–63.

BIESENBACH-LUCAS, SIGRUN. 1987. The use of relative markers in modern American English. In Denning et al., 13–21.

BOCK, J. KATHRYN. 1986. Syntactic persistence in language production. Cognitive Psychology 18.355–87.

BRANIGAN, HOLLY P.; MARTIN J. PICKERING; and ALEXANDRA A. CLELAND. 2000. Syntactic co-ordination in dialogue. Cognition 75.B13–25.

CARTER, DAVID. 1987. Interpreting anaphors in natural language texts. Chichester, UK: Ellis Harwood Ltd.

[22] It might be objected here that the requirement of RP/comp in subject relatives does not apply to all dialects of English. In a number of dialects, including African American Vernacular English (AAVE) and some variants of British English, RP/comp is optional in subject RCs (Erdmann 1980, Martin & Wolfram 1998). This might seem problematic from an ambiguity-avoidance viewpoint. We should note, however, that in both colloquial British English and AAVE, this usage appears to be quite strongly constrained. With regard to British English, Erdmann argues that a null RP in subject RCs occurs only in existential constructions (e.g. *There was someone wanted to see you*) and certain other constructions where the matrix NP is in object or predicate position (*We had a client owned a grocery shop*). It seems significant that, in the cases Erdmann mentions, the NP containing the relative clause never occurs in SUBJECT position, as this is the position where the risk of ambiguity would be greatest. In a construction like *There was someone wanted to see you*, it should be clear by the time the relative clause verb arrives that it is not the main verb. Similarly, Martin and Wolfram note that null-RP subject relative clauses in AAVE are especially associated with predicate phrases. Thus an ambiguity-avoidance account might be worth exploring as an explanation of RP/comp deletion in subject RCs. (For an alternative account of these phenomena, see Doherty 1993.)

DANCHEV, ANDREI. 1991. Language change typology and some aspects of the SVO development in English. Historical English syntax, ed. by Dieter Kastovsky, 103–24. Berlin: Mouton de Gruyter.

DENNING, KEITH; SHARON INKELAS; FAYE C. MCNAIR-KNOX; and JOHN RICKFORD (eds.) 1987. Variation in language: NWAV-XV at Stanford. Stanford, CA: Dept. of Linguistics, Stanford University.

DOHERTY, CATHAL. 1993. The syntax of subject contact relatives. Papers from the 29th regional meeting of the Chicago Linguistic Society, 155–69. Chicago: Chicago Linguistic Society.

ELSNESS, JOHAN. 1984. *That* or zero? A look at the choice of object clause connective in a corpus of American English. English Studies 65.519–33.

ERDMANN, PETER. 1980. On the history of subject contact-clauses in English. Folia Linguistica Historica 1.139–70.

FERREIRA, VICTOR S., and GARY S. DELL. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology 40.296–340.

FRAZIER, LYN. 1985. Syntactic complexity. Natural language parsing: Psychological, computational, and theoretical perspectives, ed. by David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, 129–89. Cambridge: Cambridge University Press.

GUY, GREGORY R. 1997. Competence, performance, and the generative grammar of variation. Variation, change, and phonological theory, ed. by Frans Hinskens, Roeland Van Hout, and W. Leo Wetzels, 125–43. Amsterdam: John Benjamins.

GUY, GREGORY R., and ROBERT BAYLEY. 1995. On the choice of relative pronouns in English. American Speech 70.148–62.

HANKAMER, JORGE. 1973. Unacceptable ambiguity. Linguistic Inquiry 4.17–68.

HARRIS, MARTIN B. 1978. The interrelationship between phonological and grammatical change. Recent developments in historical phonology, ed. by Jacek Fisiak. The Hague: Mouton.

HAWKINS, JOHN A. 1994. A performance theory of order and constituency. Cambridge: Cambridge University Press.

HIRST, GRAEME. 1981. Anaphora in natural language understanding: A survey. Berlin: Springer-Verlag.

KIKAI, AKIO; MARY SCHLEPPEGRELL; and SALI TAGLIAMONTE. 1987. The influence of syntactic position on relativization strategies. In Denning et al., 266–77.

KIPARSKY, PAUL. 1982. Explanation in phonology. Dordrecht: Foris.

KUNO, SUSUMU. 1974. The position of relative clauses and conjunctions. Linguistic Inquiry 5.117–36.

LABOV, WILLIAM. 1994. Principles of linguistic change: Internal factors. Oxford: Blackwell.

MARCUS, MITCHELL P.; BEATRICE SANTORINI; and MARY ANN MARCINKIEWICZ. 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19.313–30.

MARTIN, STEFAN, and WALT WOLFRAM. 1998. The sentence in African-American Vernacular English. African-American English: History, structure, and use, ed. by Salikoko S. Mufwene, John R. Rickford, Guy Bailey, and John Baugh, 11–36. London: Routledge.

MITCHELL, DON L. 1994. Sentence parsing. Handbook of psycholinguistics, ed. by Morton Ann Gernsbacher, 375–410. San Diego: Academic Press.

PYLES, THOMAS. 1971. The origins and development of the English language. New York: Harcourt Brace Jovanovich.

QUIRK, RANDOLPH; SIDNEY GREENBAUM; GEOFFREY LEECH; and JAN SVARTVIK. 1985. A comprehensive grammar of the English language. New York: Longman.

SMITH, JEREMY. 1996. An historical study of English: Function, form, and change. London: Routledge.

WASOW, THOMAS, and JENNIFER ARNOLD. 2003. Post-verbal constituent ordering in English. Determinants of grammatical variation in English, ed. by Günter Rohdenburg and Britta Mondorf, 119–54. Berlin: Mouton de Gruyter.

26 Gibbs St.
Rochester, NY 14607
[dtemp@theory.esm.rochester.edu]

[Received 26 March 2002;
revision received 3 December 2002;
accepted 3 March 2003]