Journal of Psycholinguistic Research, Vol. 25, No. 1, 1996

The Relationship Between the Frequency and the Processing Complexity of Linguistic Structure

Edward Gibson,[1] Carson T. Schütze,[2] and Ariel Salomon[1]

In this paper the relative frequencies of the possible resolutions of ambiguities involving noun phrase attachment sites are compared to the results of off-line psycholinguistic measurements of syntactic complexity. A lack of correlation between the two is observed. It is therefore argued that the comprehension system is distinct from what is driving the frequencies in the corpora. A production heuristic separate from the comprehension system is proposed to account for the observed frequencies.

INTRODUCTION

It has recently been proposed by Mitchell and colleagues (Mitchell, 1994; Mitchell & Cuetos, 1991; Mitchell, Cuetos & Corley, 1992; Mitchell, Cue-

We would like to thank Susanne Tunstall for providing linguistic judgments in the corpus analyses reported in Sec. 3.3. We also wish to thank the following people for helpful, comments and discussions on earlier presentations of this paper: the audience at the Eighth Human Sentence Processing Conference, in Tucson Arizona, Maria Babyonyshev, Tom Bever, Martin Corley, Danny Fox, Don Mitchell, Edson Miyamoto, Janet Nicol, Colin Phillips, Gregg Solomon, Michael Spivey-Knowlton, James Thomas, and Tom Wasow. All remaining errors are our own. The second author was supported by the Research Training Grant "Language: Acquisition and Computation" awarded by the National Science Foundation (U.S.) to the Massachusetts Institute of Technology (DIR 9113607), and by a fellowship from the Social Sciences and Humanities Research Council of Canada.

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

[2] Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

[3] Address all correspondence to Edward Gibson, gibson@psyche.mit.edu.

0090-6905/96/0100-0059$09.50/0

tos, Corley & Brysbaert, 1995) that initial parsing preferences in syntactically ambiguous structures are determined by people's exposure to similar structures in the past. Under this framework—the *tuning* framework—people are assumed to tabulate the resolutions of ambiguities as the ambiguities are encountered, with the result that the most frequently occurring resolution of a given ambiguity is the resolution that people tend to prefer. This framework was initially proposed in order to explain parsing preference differences between Spanish and English in relative clause (RC) attachment ambiguities such as the following:

(1) a. El periodista entrevisto a [NP_1 la hija del [NP_2 coronel]] [CP que tuvo el accidente]
b. The journalist interviewed [NP_1 the daughter of [NP_2 the colonel]] [CP who had had the accident]

In examples like these, Spanish speakers prefer attachment to the first (high) noun phrase (NP) site, while English speakers prefer attachment to the second (low) NP site inside the prepositional phrase (PP) (Clifton, 1988; Cuetos & Mitchell, 1988; Mitchell & Cuetos, 1991). To account for this difference, Mitchell and colleagues proposed the tuning framework. Thus they hypothesized that the reason for the difference between the English and Spanish preferences is that there is a difference in the relative frequencies of the resolutions of similar ambiguities in the input English and Spanish speakers are exposed to. (See Carreiras & Clifton, 1993; De Vincenzi & Job, 1993; Gibson, Pearlmutter, Canseco-Gonzalez, & Hickok, in press; Gilboy, Sopena, Clifton, & Frazier, 1995, for some alternative hypotheses to explain the parsing preference differences between the two languages.)

The most direct way of testing this hypothesis is to analyze large corpora of naturally occurring texts to see if this ambiguity has different frequencies of resolution in the two languages. Mitchell, Cuetos, and Corley (1992) reported that this seems to be the case in a small-scale study of Spanish and English corpora. In their preliminary analyses of instances of two-site RC attachments, 60% of the RCs in the Spanish examples attached to the high site, while only 38% of the RCs in the English examples attached to the high site, as expected under the tuning hypothesis (numbers from Cuetos, Mitchell, & Corley, in press). However, although this is promising evidence in support of the tuning hypothesis, Cuetos et al. make it clear that the Mitchell et al. (1992) study was only preliminary, and that much further evaluation is needed, not only for the two-NP-site ambiguities, but also for other kinds of ambiguities.

Although the steps for testing the tuning hypothesis have been clear, progress has been slow because obtaining ambiguity-resolution frequency counts has proved to be very labor-intensive. To obtain the relevant struc-

tural frequencies, it is necessary (1) to have a large corpus of naturally occurring texts, and (2) to be able to locate all the relevant components of these texts which contain the target structures. Large unprocessed English corpora are now quite abundant (see, e.g., the list of corpora available from the Linguistics Data Consortium at the University of Pennsylvania), but corpora in languages other than English are less easily available. Furthermore, corpora from all languages (including English) in which target syntactic structures can be accessed are less available. Optimally, we would like an automatic procedure to accurately parse the corpora, so that the resulting parses could be searched for the desired constructs.[4] However, the best currently existing computational parsers do not have the required accuracy to make this goal achievable. Thus the only way at present to parse unrestricted texts with useful accuracy is to parse the texts by hand, perhaps with some initial automatic parsing stage. This is what has been done at the University of Pennsylvania Treebank project (Marcus, Santorini, & Marcinkiewicz, 1993), where over two million words of naturally occurring American English texts have been hand-parsed.

In order to further test the tuning hypothesis, this paper reports frequency counts from these corpora as well as experimental results from a comprehension complexity experiment on the resolution of the ambiguity in Fig. 1, which involves three NP attachment sites.

[4] For the purposes discussed here, it might be possible to develop heuristics which find all (or most) of the target ambiguous structures in unrestricted text, and then parse these by hand to determine the resolution frequencies. This is the approach taken in the CORSET project (Corley & Corley, 1995). For such an approach to be successful, it will be necessary to evaluate it against a fully parsed corpus to determine that the sample of structures obtained by the heuristics is not biased in some ways.

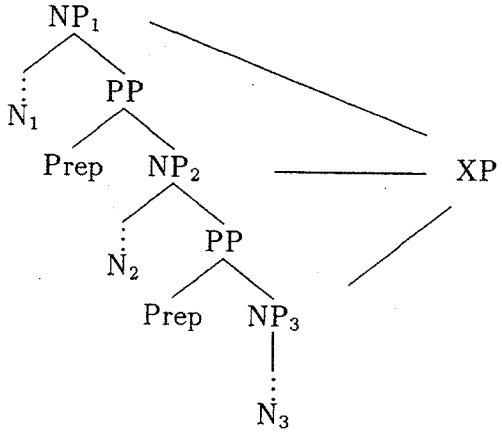

Fig. 1. Ambiguous attachment of a phrase (XP) to three prospective noun phrase (NP) sites. N = noun; PP = prepositional phrase; Prep = preposition.

The three-NP-site ambiguity is one in which there is strong processing evidence for a preference ordering among the prospective attachment sites. Gibson et al. (in press) provide both off-line and on-line evidence that the attachment ranking in such an ambiguity when the attaching item XP is an RC consists of low attachment (NP_3) first, followed by high attachment (NP_1), with attachment to the middle site (NP_2) hardest of all. Consider the examples in (2), in which the attaching RC agrees in number with only one of the three prospective heads:

(2) a. Low: The lamps near the paintings of the house that was damaged in the flood
b. Middle: The lamps near the painting of the houses that was damaged in the flood
c. High: The lamp near the paintings of the houses that was damaged in the flood

Gibson et al. found that subjects read the disambiguating region of the RC fastest when it agreed with the low attachment site, second fastest when it agreed with the high site, and slowest when it agreed with the middle site. On-line and off-line grammaticality judgments also confirmed this preference ordering.

Furthermore, the same preference ordering of (low, high, middle) is also observed in Spanish, in spite of the evidence that Spanish speakers prefer high attachment in two-site cases (Cuetos & Mitchell, 1988; Mitchell & Cuetos, 1991). In order to account for these facts, Gibson et al. (in press) proposed that the parser is governed by two interacting constraints (among others):

(3) Recency preference (Gibson, 1991; cf. late closure, Frazier, 1978, 1987; Frazier & Rayner, 1982; cf. right association, Kimball, 1973): Preferentially attach structures for incoming lexical items to structures built more recently.
(4) Predicate proximity: Attach as close as possible to the head of a predicate phrase.

Recency preference is assumed to follow from a short-term memory constraint that causes attachment sites to decay in their activation over time, so that less recent sites are less preferred than more recent sites. Following evidence from the short-term memory literature (see, e.g., Anderson, 1980, 1983; McClelland & Rumelhart, 1981), it is assumed that recency preference decays according to an exponential function, approaching an asymptote in the limit. Thus recency preference by itself predicts a monotonic preference ordering of (NP_3, NP_2, NP_1), with the difference between the costs of at-

tachment to NP_3 and NP_2 more than the difference between the costs of attachment to NP_2 and NP_1.

According to predicate proximity, the attachment which is structurally closest (i.e., in terms of the number of tree nodes) to the head of a predicate phrase (verb phrase) is preferred over all other attachments. If the head NP, NP_1, of a three-NP sequence is attached as the argument of a verb, it will therefore be the preferred attachment site according to predicate proximity. The difference in preferences in two-site attachments between English and Spanish is hypothesized to follow from a parameterization in the cost associated with violating predicate proximity across the two languages such that the cost associated with predicate proximity is high in Spanish relative to English (see Gibson et al., in press, for motivation).

The lack of a difference between the two languages with respect to the three-site preferences emerges because of the decay function associated with recency violations. When there are three sites, attachment to the high NP site incurs more recency cost than if there are only two sites, while the cost associated with predicate proximity is assumed to be the same for each of the non-predicate-proximate sites.[5] Thus while the high site is less costly than the low site in Spanish two-site attachments (because violating predicate proximity is more costly than violating recency in Spanish), the high site is more costly than the low site in a Spanish three-site ambiguity because of the additional recency violations. (See Gibson et al., in press, for more details.)

Let us now consider how the tuning hypothesis might explain the same preferences. First of all it should be noted that the relative ordering of the high and low sites in three-site cases cannot follow from tuning on two-site preferences, because this preference ordering switches from a high preference to a low preference in two- versus three-site ambiguities in Spanish. Similarly, while the relative ordering between the high and low sites appears to be the same in English with respect to two- and three-site ambiguities, intuitions suggest that the preference for the low site appears to be stronger in three-site ambiguities than it is in two-site ambiguities. Hence in order to account for the preference orderings and their relative strengths, the tuning hypothesis must distinguish two- and three-site ambiguities.[6] In particular,

[5] The decay function associated with cost increases for predicate proximity does not actually have to be an all-or-nothing step function, as is assumed here. The rate of decay only needs to be more rapid than that associated with recency. One of the simplest such functions is the all-or-nothing step function, so that is what is initially assumed.

[6] It is also possible that the relevant difference between the two-site and three-site examples that have been studied so far is not the number of sites, but is the distance in words or morphemes between the high and low site. Whatever this difference turns out to be, the tuning framework will need to keep separate counts for the relevant differences in ambiguity.

the tuning framework will need to set preferences for three-site ambiguities based on three-site input, not two-site input. Thus to see if the tuning framework is workable, we examine the frequencies of resolutions in favor of each of the attachments in three-site ambiguities in large corpora.

Even if the comprehension complexity of an ambiguity resolution correlates with its frequency in the input, this does not necessarily mean that the tuning hypothesis is correct: A correlation between frequency and comprehension complexity is consistent with other possibilities. One possibility is that linguistic comprehension complexity (as quantified by an independently motivated linguistic comprehension theory) is driving the production frequencies, so that structures which are harder to understand are produced less often. A related possibility is derived from the fact that the corpora being analyzed are corpora that have been edited at some stage (e.g., newspapers, magazines, books, etc.) as opposed to unedited corpora (e.g., naturally occurring dialogues). Because these are edited corpora, it could be that the correlation between the resolution frequencies and their processing complexity is due to the editing process: Presumably the editing is intended (in part) to make the texts easier to understand. If editing is driven in part by the linguistic comprehension mechanism, and this mechanism is not driven by frequency of input alone, then a correlation between frequency and comprehension complexity might exist without implying the existence of a tuning framework. Hence finding a correlation between frequency and comprehension complexity does not imply that the tuning framework is correct.

Of course, it is also possible that frequency and comprehension complexity do not correlate for a given construction, even in edited corpora. This may happen if (1) the sentence comprehension mechanism is not guided by the frequencies of ambiguity resolution, and (2) the sentence production mechanism is not driven solely by comprehension difficulty. However, the lack of correlation between the frequency and comprehension complexity of a particular construction is actually not sufficient to rule out the tuning hypothesis, even for that construction. It could be that the "grain size" of the ambiguity being analyzed is not one that the parser specifically considers. Rather, the tuning parser may be considering either more general or more specific instances of the construction in question, or perhaps the tuning parser is tuning in a separate dimension for this construction, so that the observed divergences in frequency and complexity may be artifacts of examining a grain size different from that which is considered by the parser (see, e.g., Gibson & Pearlmutter, 1994, and Spivey-Knowlton & Sedivy, 1995, and the discussion of these papers in the second section). Thus interpreting a lack of correlation is difficult. In order to argue that more than resolution frequency is driving the comprehension mechanism for the con-

structions in question, it is necessary to look at finer grains and coarser grains, and show that each of these is unworkable.

With respect to the three-NP-site ambiguities considered here, we have argued above that the coarse grain of tuning three-site preferences based on two-site inputs is not workable because of the differences in preference orderings and their relative strengths in two- and three-site ambiguities. Hence we have established a lower bound on the grain-size in terms of coarseness. It is also possible that tuning takes place across all kinds of attaching categories, such as relative clauses, prepositional phrases, VP modifiers, adjectival modifiers, etc. Or it might be that each kind of attaching category is tuned separately. In order to test the tuning hypothesis with respect to the three-NP-site ambiguity, it is necessary to show that the same grain-size works for all of these different attaching categories. On the other hand, to demonstrate that the tuning theory is not workable for a given construction, it is necessary to show that there is no grain-size that can plausibly account for the preferences. This is the type of result that we are led to here: No matter what the grain-size in the three-NP attachment ambiguities, the frequency/complexity correlation does not occur for certain attaching categories.

Previous related work in this area is described in the second section. The third section describes the new evaluation of the frequency/complexity hypothesis. This evaluation consists of frequency counts together with results from a new comprehension complexity experiment for three-NP-site ambiguities. A lack of correlation between the comprehension complexity and the frequencies is observed. In the third section we also analyze the variable grain-size solution to the problem, and it turns out that no matter what grain is selected, the lack of correlation is still observed. In the fourth section we propose a theory of these frequencies. Conclusions are given in the final section.

PREVIOUS WORK

In one of the first major studies of the frequencies of syntactic ambiguity resolution, Hindle and Rooth (1993) observed that, in PP attachment ambiguities involving a preceding verb and an NP, the PP attached to the NP 67% of the time in a hand-parsed sample of 880 randomly selected instances of the ambiguity from the 1989 *Wall Street Journal.* If the tuning hypothesis is correct and tuning takes place with the grain size of these categories, then the preferred attachment site in such ambiguities should be the NP. However, as observed in Mitchell and Cuetos (1991), this prediction

contrasts with Rayner, Carlson, and Frazier's (1983) experiments, which found that the attachment to the verb is the one that people preferred over their experimental materials (see also Clifton, Speer, & Abney, 1991; cf. Taraban & McClelland, 1988).

For the tuning framework to be consistent with this result, it must therefore be the case that the human parser does not tune ambiguity resolutions at this grain size. That is, the human parser must be tuning its preferences based on more coarse-grained or more fine-grained categories in order to account for the experimental result. The finer-grain hypothesis is consistent with a more recent corpus analysis of PP attachments performed by Spivey-Knowlton and Sedivy (1995). Spivey-Knowlton and Sedivy obtained the corpus counts shown in Table I from half of the Brown corpus of English text (Kučera & Francis, 1967) for V–NP–PP ambiguities in which the PP is headed by *with*. They found that attachments to the VP were more prevalent overall, but that the distributions varied according to (1) the type of the verb involved, either action verb or nonaction verb (e.g., a psychological or perception verb or *have* or *be*), and (2) whether the NP attachment site was definite or indefinite.

Spivey-Knowlton and Sedivy's (1995) first self-paced reading experiment demonstrated that, when the verb is an action verb and the NP site is definite, the preference is to attach to the VP, consistent with the corpus numbers. This experiment also demonstrated a weaker preference to attach to an action-verb VP for indefinite NP cases. Although not predicted by the corpus numbers, this second result is consistent with them because the corpus numbers are not significantly different in this case. Spivey-Knowlton and Sedivy's second reading experiment tested psychological and perception verbs. An NP attachment preference was revealed for the indefinite NP case, as predicted by the corpus numbers. This experiment also showed a VP

Table I. Frequencies from Half of the Brown Corpus of Prepositional Phrases (PPs) headed by *with* Attaching to a Preceding Noun Phrase (NP) or Verb Phrase (VP) (Spivey-Knowlton & Sedivy, 1995)

	NP attached	VP attached
Action verbs		
Definite NP	0	31
Indefinite NP	8	9
Total	8	40
Psychological and perception verbs		
Definite NP	4	3
Indefinite NP	10	1
Total	14	4

attachment preference for the definite NP case, which, although not predicted by the corpus numbers, is consistent with them, again because the numbers are not significantly different. Spivey-Knowlton and Sedivy's corpus analyses and reading experiments are therefore consistent with the general tuning framework: The attachment preferences for a V–NP–PP ambiguity are tuned according to the class of the verb involved (e.g., perception verb, psychological verb, action verb) and the definiteness of the NP attachment site. Of course, it should be kept in mind for all of these examples that it could also be that it is something in the syntax and semantics of the syntactic categories and lexical items in question that leads to the comprehension complexities and frequencies in the input, so that the fact that the data are currently consistent with the tuning hypothesis does not imply that tuning is the real explanation for these results.

Other studies have also shown correlations between the comprehension complexities of lexical ambiguity resolution and the frequencies of the ambiguities in the input (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994, for the reduced relative/main verb ambiguity; MacDonald, 1993, for noun/verb category ambiguities, among others). However, to the extent that there are general syntactic preferences which operate after lexical effects have been factored out, then lexical tuning will be insufficient and higher-level category tuning will also be necessary. Evidence for the insufficiency of lexical frequency effects alone in accounting for syntactic preferences is provided by Juliano and Tanenhaus (1994) and Merlo (1994) with respect to the S/NP-complement ambiguity, and by Mitchell et al. (1995) with respect to the RC attachment ambiguities involving two NP sites.

The three-NP-sites attachment ambiguity in Fig. 1 provides another instance of an ambiguity that is probably not a lexical one. Previous work on the resolution of this ambiguity has examined both PP attachments and RC attachments in English (Gibson & Loomis, 1994; Gibson & Pearlmutter, 1994). As discussed in the Introduction, the preference ordering among the three sites for RC attachment is (low, high, middle) (or equivalently, NP_3, NP_1, NP_2). It is hypothesized by Gibson and Pearlmutter based on intuitive judgments that the same preference ordering holds for all modifier attachments (such as PPs) in an ambiguity like that in Fig. 1.

The frequency evidence put forward by Gibson and Pearlmutter (1994) came from their analysis of as many PP attachments matching the configuration in Fig. 1 as they could locate in the Brown corpus. The evidence presented by Gibson and Loomis (1994) relevant to the three-NP-site ambiguities came from their analysis of as many PP and RC attachments matching Fig. 1 as they could find in the parsed *Wall Street Journal* (WSJ) corpus of the University of Pennsylvania Treebank (Marcus et al., 1993). These initial corpus counts are given in Tables II and III, respectively.

Table II. Brown Corpus Frequencies of Attachment of a Prepositional Phrase (PP) to One of the Three Preceding Noun Phrase (NP) Attachment Sites in the Syntactic Configuration from Fig. 1 (Gibson & Pearlmutter, 1994)

	Attachment site		
	NP_1	NP_2	NP_3
Number of PP tokens	62	63	204

Table III. Penn Treebank *Wall Street Journal* Corpus Frequencies of Attachment of a Prepositional Phrase (PP) or Relative Clause (RC) to One of the Three Preceding Noun Phrase (NP) Attachment Sites in the Syntactic Configuration from Fig. 1 (Gibson & Loomis, 1994)

	Attachment site		
	NP_1	NP_2	NP_3
Number of PP tokens	92	144	376
Number of RC tokens	22	15	150

Although the RC attachment frequencies are numerically ranked in the same sequence as the comprehension complexity ordering (if not significantly so), the PP attachments are not. If all attaching categories have the same attachment preferences as RCs, so that high attachments are easier to comprehend than middle attachments for PPs as well as RCs, then tuning cannot be taking place at this grain size. In particular, middle attachments are equally frequent according to Gibson and Pearlmutter's (1994) Brown corpus numbers, and are more frequent according to Gibson and Loomis's (1994) WSJ corpus numbers.

It was argued in the introductory section that tuning cannot be taking place at a coarser grain of analysis for RC attachments (e.g., two sites). Assuming that the tuning mechanism applies with the same grain size across all ambiguities, tuning cannot be taking place at a coarser grain of analysis for PP attachments either.[7] So finer grains need to be considered. In fact, additional analyses provided in both studies suggest that a finer grain size is consistent with the tuning hypothesis. In particular, if only examples are considered in which there is no independent reason why one of the sites is

[7] It is also possible that the grain size of tuning depends on the number of instances that are encountered, such that more frequently occurring ambiguities are tuned at a smaller grain size. This hypothesis still rules out the possibility that PP attachments are tuned on input from two NP sites, because they are more frequent than RC attachments of the same kind.

less likely to serve as an attachment site than the others, then the frequencies look much more like those expected by the tuning hypothesis. The following kinds of cases were therefore excluded from the counts:

1. Items in which one of the sites which was tagged as an NP was part of an idiomatic expression, such as *in spite of* or *in connection with.*
2. Items in which any of the prospective attachment sites was part of a complex proper name that includes at least one prepositional phrase, such as *the State University of New York* and *Committee for the Scientific Branch of the Paranormal.*
3. Items in which one of the NP sites was a quantifier such as *one of* or *some of,* which are less likely attachment sites.
4. Items in which the presence of intervening punctuation such as quotation marks, commas, or dashes made at least one of the attachment sites less likely.
5. Items in which the attaching phrase (PP or clause) is an argument or is closely linked lexically with one of the three prospective sites. It is well known that argument attachments are generally preferred over modifier attachments. Thus only cases involving potential modifier attachments were considered here. Examples of prepositional arguments of NPs that were filtered by Gibson and Pearlmutter (1994) are given in (5); an example of a clausal complement structure that was filtered is *the fact that. . . .*

   (5) a. Low: the lack of scientific unanimity on the effects [of radiation]
   b. Middle: the host of novel applications of electronics [to medical problems]
   c. High: the relation of the figure of the dancer [to light and color]

Finally, items in which the attaching item could attach in more than one location without distinguishable differences in meaning were also filtered, because they could be associated with multiple sites (Hindle & Rooth, 1993; Hobbs & Bear, 1990).

Most of this filtering process is straightforward and objective, but two of the steps are labor-intensive and subjective. The difficult steps involve making judgments on (1) whether an attaching item is an argument or a modifier (whether it is lexically preferred or not) and (2) whether two different attachments of the same phrase result in the same meaning. These steps were performed independently by the two researchers in each case, but because they involve subjective judgments, their validity is somewhat questionable.

In any case, once these items were removed, only unambiguous modifier attachments remained, whose frequencies now roughly mirror the comprehension complexity observed in the processing experiments (see Tables IV and V). Tuning is therefore consistent with these results.

Table IV. Brown Corpus Unambiguous Modifier Prepositional Phrase (PP) Attachments (NP = Noun Phrase) (Gibson & Pearlmutter, 1994)

	Attachment site		
	NP_1	NP_2	NP_3
Number of PP tokens	27	10	68

Table V. *Wall Street Journal* Corpus Unambiguous Modifier Prepositional Phrase (PP), Relative Clause (RC) Attachments (NP = Noun Phrase) (Gibson & Loomis, 1994)

	Attachment site		
	NP_1	NP_2	NP_3
Number of PP tokens	20	10	68
Number of RC tokens	7	2	79

However, there are reasons to be cautious about these results. Most importantly, the programs that searched for matching patterns are based on heuristics, and hence do not find all the instances of matching attachments. If matching instances are being missed in systematic ways, then these results are not necessarily meaningful. In fact, our results below show that these counts highly underestimate the number of matching attachments in the corpora in question. In particular, the true number of matching attachments is more than double what these studies reported for each of the attachment sites, with a greater number of middle attachments missed than high attachments.

CONJUNCTION ATTACHMENTS: CORPUS-BASED AND COMPREHENSION COMPLEXITY EVIDENCE

Although the filtered analyses presented by Gibson and Pearlmutter (1994) and Gibson and Loomis (1994) are consistent with the tuning hypothesis, it would be desirable to get independent support for these counts because of the subjective nature of many of the judgments involved, particularly the lexical preference judgments, and the identical meaning judgments. Given the difficulties associated with analyzing PP and RC attachment ambiguities, together with the time-consuming nature of this task, we analyzed another attachment ambiguity to compare frequency and comprehension complexity: NP conjunctions. The particular construction type that

is examined here involves a conjunction of noun phrases within the three-NP-site right-branching structure discussed above, as illustrated in the tree in Fig. 2.

The advantage of analyzing NP conjunction attachments as opposed to RC attachments or PP attachments is that much less filtering by hand is needed for the NP conjunction attachments, because (1) few items have lexical requirements for a conjoined element, so judgments of argumenthood are much less frequent; and (2) attachments to the three different sites almost universally result in differences in meaning which are easily distinguishable, so that much less hand filtering is done in this step also. The corpus analyses can therefore be more automatic, with many fewer human judgment calls.

Corpus Analyses

The parsed Brown and *Wall Street Journal* corpora from the Penn Treebank were analyzed, searching for the pattern in Fig. 2 below. Automatic searches using the *tgrep* utilities provided with the Penn Treebank yielded the pattern of frequencies in Table VI.

These frequencies pattern like the unfiltered PP attachment frequencies from Gibson and Loomis (1994): Low attachments are the most common [vs. middle: Brown: $\chi^2(1) = 133.6$, $p < .001$; WSJ: $\chi^2(1) = 257.4$, $p < .001$], followed by middle attachments, with high attachments least frequent [vs. middle: $\chi^2(1) = 16.79$, $p < .001$; WSJ: $\chi^2(1) = 2.31$, $p = .12$].[8] The number of low attachments is so much larger than either of the other two

[8] All χ^2 tests with exactly one degree of freedom in this paper have Yates' correction for continuity applied (Hays, 1988).

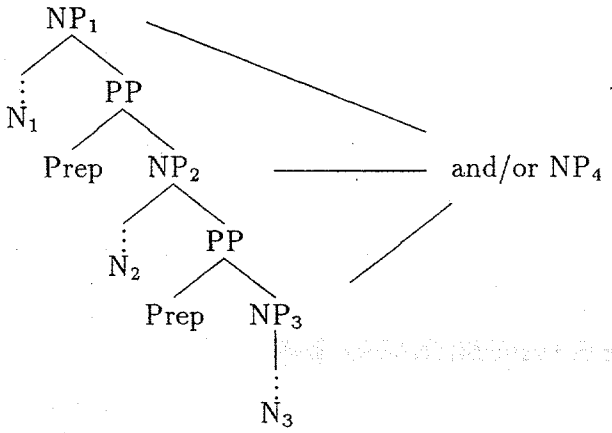

Fig. 2. Ambiguous attachment of a conjoined noun phrase (NP) to three prospective NP sites. N = noun; PP = prepositional phrase; Prep = preposition.

Table VI. Unfiltered Frequencies of Noun Phrases (NPs) Conjoined to NPs with three NP sites, as in Fig. 2

	Corpus	
Attachment site	Brown	WSJ[a]
NP_1 (high)	54	68
NP_2 (middle)	107	88
NP_3 (low)	357	467

[a] WSJ = *Wall Street Journal.*

attachment frequencies that this difference is unlikely to change after hand filtering. However, the ordering of middle with respect to high might be due to extraneous peculiarities of the examples involved in the high and middle attachment frequencies, as was the case for the PP attachments. The outputs of the high and middle searches were therefore filtered by hand to eliminate examples that were biased in various ways, in much the same way as done by Gibson and Pearlmutter (1994) and Gibson and Loomis (1994) with respect to PP and RC attachments. As noted above, this filtering is considerably more straightforward than for the PP and RC attachments, because many fewer subjective lexical preference and semantic equivalence judgments are necessary. The types of items that were filtered are as follows:

1. Items containing idioms
2. Items containing punctuation: commas, quotation marks, dashes, etc.
3. Items containing names that include one of the PPs, e.g., *United States of America*, which bias against conjoining within them, unless the proper name includes a conjunction, in which case the bias is to complete the proper name
4. Items that include *between* or *both*, which create a bias to take a following *and* matching at the same level [see (6)]
5. Items that already contain a conjunction
6. Items in which one of the nouns requires a semantically plural complement, e.g., *combination*, which are biased to attach *and* directly under their scope in order to fulfill the plurality requirement

(6) [NP_1 the relation between [NP_2 the Protestant movement in [NP_3 this country] and [NP_4 the development of a social religion]]]

In (6), *relation between* requires more than one entity in its complement, and since the first NP is singular (*movement*), high attachment is ruled out.

The frequencies of the remaining examples are as in Table VII. Examples of the remaining high and middle attachments are given in (7):

(7) a. Middle: [NP_1 calculator-toting, socially awkward individuals with [NP_2 shirt-pocket liners for [NP_3 their pencils and [NP_4 a preoccupation with computers and matters numerical
b. High: [NP_1 the brilliant clash of [NP_2 styles of [NP_3 narration and [NP_4 the even more brilliant way that they have been tied together into a large metaphor for literature and its function in society

Middle attachments are still more frequent than high attachments, marginally more so in the Brown corpus [$\chi^2(1) = 19.64$, $p < .09$], nonsignificantly so in the WSJ corpus. Hence the tuning hypothesis predicts that, in a comprehension experiment, people should prefer low attachments first, followed by middle attachments and high attachments about equally preferred. If there is any preference between the latter two, the tuning hypothesis predicts that it should be towards the middle attachment site.

The tuning theory therefore makes the opposite prediction of the recency/predicate proximity theory discussed earlier with respect to the high and middle sites in these examples. The recency/predicate proximity theory predicts that high attachments should be easier than middle attachments, for the same reasons that RCs and PPs are preferentially attached high.

Experiment 1: Comprehension Complexity Evidence

In order to test the predictions of the theories, an acceptability-judgment questionnaire was conducted to assess the processing complexity of three-NP-site conjunction attachments, using examples like (8):

(8) The salesman ignored the customer with the child with the dirty face and
a. the wet diaper. [low]

Table VII. Filtered Frequencies of Noun Phrases (NPs) Conjoined to NPs with three NP sites

	Corpus	
Attachment site	Brown	WSJ[a]
NP_1 (high)	22	29
NP_2 (middle)	36	37

[a] WSJ = *Wall Street Journal.*

b. the one with the wet diaper. [middle]
c. the one with the baby with the wet diaper. [high]

The attachment site for the conjunction *and* is disambiguated to one of the three NP sites in two ways.[9] First, each of the completions was pragmatically disambiguated. In the low-attached (a) completion, high attachment is unlikely because salesmen do not usually ignore wet diapers; the middle attachment reading is also disfavored, because customers do not usually have wet diapers. In the middle-attached (b) completion, a low attachment would require conjunction of *the dirty face* with *the one with the wet diaper*. Because the word *one* requires a contrasting modifier, and because the only available contrasting modifier is the adjective *dirty* (modifying *face*), the interpretation of this conjoined NP associates the PP *with the wet diaper* with the noun *face*. This attachment is pragmatically ruled out, because faces do not normally have wet diapers. The low attachment of the (c) completion is pragmatically ruled out in a similar way, because faces do not usually have babies (with wet diapers).[10] After *the one*, the (b) and (c) completions are also disambiguated pragmatically. Because customers do not usually have wet diapers, the (b) completion is unlikely to be conjoined at the high site. Finally, because children do not usually have babies, the (c) completion is unlikely to be conjoined at the middle site.

The second way that the completions were biased toward an attachment site was by manipulation of the second conjunct so that it was maximally parallel to its conjoining element in terms of length and structure. In particular, the conjoined NPs in the complete structures for each of the three versions of (8) contain the same number of PPs (zero, one, or two, for low, middle, and high conjunction, respectively) and are right-branching. Assuming there is a general preference for conjoined constituents to be maximally parallel, the intended attachments will be the preferred ones. This also has the desirable consequence that degree of parallelism is controlled for across

[9] When the conjunction itself is first processed, higher attachment sites are possible, including VP- and S-level conjunction. The former is ruled out by the word following the conjunction. The latter remains possible in principle until the end of the sentence; however, Frazier (1978) has shown that S-level conjunction is dispreferred relative to NP conjunction, so we assume that subjects were not pursuing this possibility.

[10] In fact, not all of our examples included adjectival modifiers for the low attachment site. For these examples (the minority), low attachment for the (b) and (c) completions is grammatically ruled out because there is no contrasting modifier at this site to allow attachment.

the three completion conditions: They are all as parallel as they can be structurally.[11]

The tuning theory predicts that the comprehension preferences should roughly mimic the frequency distributions found in the corpora: The low attachment should be the easiest to process, followed by the middle and the high site. On the other hand, the processing theory which includes the principles of recency preference and predicate proximity predicts that the attachment preferences in these examples should be much the same as in the RC attachments: The low attachment should be easiest to process, followed by the high attachment, with the middle attachment the worst of the three.[12]

An extension of the discourse-based theory of Crain and Steedman (1985) and Altmann and Steedman (1988) might also predict the high-attachment preference over the middle attachment. Under this theory, the preferred structure in an ambiguous linguistic input is the one that requires the minimal discourse structure, as indexed in part by the number of discourse objects that are necessitated. Much of the empirical evidence for this theory is based on the claim that an unmodified definite NP pragmatically presupposes the existence of a unique referent in the discourse. Furthermore, following a Gricean maxim, if a definite NP is modified, then a set of similar objects is presupposed to exist, of which only one has the modificational

[11] A reviewer raises the possibility that the preference for parallel conjoined elements might be computed on word strings before syntactic structure is computed. Given this hypothesis, the degree of parallelism between conjuncts could be confounded with attachment site in this paradigm. In particular, the high completion (8c) contains the same amount of NP material before and after the conjunction, while in the middle and low completions there is less material after than before the conjunction, so that the high attachment is most "string-parallel" and the low attachment is least "string-parallel." However, for such a strategy to be in operation, the parser must crucially be comparing elements *before* structuring them, and yet somehow choosing the direct object (the highest postverbal NP) as the element with which to compare the second conjunct. There is no independent evidence for a processing stage in which unstructured strings of words are operated upon. Indeed, all current processing evidence suggests that people structure input almost immediately after they encounter it, not lagging a few words behind, as would be necessary here. Thus we think it is unlikely that such a prestructure string-parallelism constraint is in operation. Furthermore, intuitions suggest that altering our paradigm to make all three completions equally long (containing two PPs) does not affect the relative difficulty of the attachments.

[12] Note that in order to control for parallelism of conjuncts, the three different completions are not matched for length: The high-attachment items are longest, followed by the middle-attachment items, with the low-attachment items shortest. If there is any difference in acceptability rating due to length, it will likely decrease acceptability with length. Hence, if anything, this confound works against the most interesting prediction of the recency/predicate proximity theory: That high attachments should be easier to process than middle attachments. To the extent that a high-attachment preference is observed relative to middle attachments, the result is therefore strengthened.

property. The evidence for this proposal includes a number of studies which have demonstrated that, out of context, definite NPs are behaviorally less likely to take restrictive modifiers than are indefinite NPs (Crain & Steedman, 1985; Spivey-Knowlton & Sedivy, 1995). An extension of this theory might predict a high-attachment bias in (8), as follows. According to this theory, a set of customers, each with one or more children, is already presupposed at the point of processing the conjunction *and* in (8); otherwise the modifiers would not have been included. Conjoining the following NP to the high site does not add to the discourse presuppositions in any way. But conjoining to either the middle or the low site adds additional presuppositional structure. If the NP conjoins to one of these sites, then the discourse theory presupposes the existence of a set of customers each with a child who has a dirty face, and only one of which has the property specified in the conjoined NP. Thus, this discourse-based theory favors high over middle and low attachments in examples like (8), with no prediction about a difference between middle and low attachments.

To investigate whether definiteness might be causing any processing preferences we might find with examples like (8), corresponding sentences with indefinite determiners were also tested, as in (9):

(9) The salesman ignored a customer with a child with a dirty face and
  a. a wet diaper. [low]
  b. one with a wet diaper. [middle]
  c. one with a baby with a wet diaper. [high]

The Crain and Steedman (1985) style discourse-based theory predicts that these three conditions should be processed more easily than their definite counterparts, because unlike the modification of definites in a null context, the modification of indefinites does not presuppose the existence of additional sets of discourse objects. The discourse-based theory makes no predictions regarding the relative complexity of the three attachment locations, however. In contrast, the recency/predicate proximity theory makes the same predictions for the indefinite cases as it does for the definite items—a preference ordering of low, high, middle—with no difference in preference ordering between definites and indefinites predicted.

Method

Subjects. Thirty-six native English speakers from MIT (primarily undergraduate students) participated, for $5.00 each.

Materials. Twenty-four complex NP items with six forms like those shown in (8) and (9) were constructed. Each item contained an initial NP followed by two PPs and a further conjoining NP. Lexical and pragmatic

constraints were used to maximize the likelihood that subjects would interpret the structure preceding the conjunction as purely right-branching, with the first PP modifying the first NP, and the second PP modifying the second NP. The items were constructed so that the conjoining NP could plausibly conjoin to only one of the three prospective NP attachment sites: low, middle, or high. Each disambiguating completion was constructed to be as plausible as possible. For the high-attachment condition, this involved ensuring that the first noun [e.g., *customer* in (8)] could take a conjoined modifier; i.e., a customer can be with two people or things as easily as with one. For the middle-attachment condition, this involved ensuring that the second noun [*child* in (8)] could take a conjoined modifier; i.e., a child can have two attributes as easily as one.

Furthermore, each of the three attachment site conditions came in two definiteness conditions: one in which all NP sites were initiated by the definite determiner *the* (the definite condition) and one in which all NP sites were initiated by the indefinite determiner *a* (the indefinite condition). Thus each item had six versions, one for each of {definite, indefinite} crossed with {low, middle, high}. The items are available from the authors upon request or by anonymous ftp to psyche.mit.edu (Internet address 18.88.0.85) in the file pub/gibson/jpr95.materials.

The 24 experimental items were combined with 56 fillers to form six lists. The fillers were of approximately the same length and complexity (number of words and constituents) as the experimental items. The experimental items were counterbalanced across the lists so that each list contained exactly one version of every item. Four practice items were also constructed to be similar to the fillers.

Procedure. The stimuli were presented in the form of a questionnaire in which subjects were asked to rate sentences on a scale from 1 (*best*) to 5 (*worst*) according to how easy or hard to understand the sentences were on the first reading. The sentences were presented 10 to a page in one pseudorandom order for each list (at least one filler separated any pair of experimental items). The order of pages was randomized for each subject.

Results

The mean ratings and standard errors of each of the experimental conditions are given in Table VIII.

Two separate 3 (Attachment Site) × 2 (Definiteness) analyses of variance (ANOVAs) were conducted, treating either subjects or items as the random factor. There was a significant main effect of attachment site [$F_1(2, 70) = 42.48$, $p < .001$; $F_2(2, 46) = 39.96$, $p < .001$], but no interaction with definiteness [$F_1(2, 70) < 1$; $F_2(2, 46) < 1$]. Collapsing across definites

Table VIII. Experiment 1 Mean Ratings and Standard Errors (in Parentheses) by Condition

Attachment Site	Definite	Indefinite	Mean
High	3.14 (0.14)	2.83 (0.14)	2.99 (0.10)
Middle	3.38 (0.14)	3.27 (0.16)	3.33 (0.11)
Low	2.43 (0.14)	2.32 (0.11)	2.38 (0.09)
Mean	2.99 (0.09)	2.81 (0.09)	2.90 (0.06)

Note: Ratings were from 1 (*easy to understand*) to 5 (*hard to understand*).

and indefinites, the low attachment site was rated better than the high site [$F_1(1, 35) = 43.67, p < .001$; $F_2(1, 23) = 40.04, p < .001$] and the high site was rated better than the middle site [$F_1(1, 35) = 10.50, p < .005$; $F_2(1, 23) = 8.547, p < .01$]. There was also a main effect of definiteness, with the indefinites being rated easier than the definites [$F_1(1, 35) = 4.826, p < .05$], although this effect fell just short of significance in the items analysis [$F_2(1, 23) = 3.125, p < .10$].

Discussion

The preference ordering among the attachment sites is as predicted by the recency/predicate proximity theory, but not as predicted by the tuning theory: Just as with RC attachments, the order of attachment preferences for a conjoining NP attaching to one of three sites in a right-branching structure is (low, high, middle), for both definite and indefinite attachment sites. This contrasts with the ordering of frequencies of these attachments that was observed in the Brown and WSJ corpora described above.

The main effect of definiteness was as predicted by the Crain and Steedman (1985) discourse theory. However, the lack of an interaction between the attachment site conditions and the definiteness conditions is not expected under this theory. So although there is some validity to the discourse theory, it is probably not relevant to the attachment site ambiguity being studied here.

The tuning theory is not yet ruled out by the observed lack of correlation between frequency and complexity, however. One possibility for an explanation of the noncorrelation is that the input from which the behavioral patterns are learned may be different from the corpora being analyzed. That is, it could be that the particular parsed corpora in the Penn Treebank are not representative of typical English text. Although this is certainly a possibility, some doubt is cast on this possibility by the fact that the frequency distributions for the constructions in question look very similar in the two

corpora [as do the frequencies for related constructions (Gibson & Loomis, 1994; Gibson & Pearlmutter, 1994)], in spite of the fact that the two corpora being analyzed are very different kinds of corpora.

A related possibility is that the tuning preferences for a given language might be based on spoken corpora rather than written corpora, so that the results from the analyses of written texts may not be directly relevant to the issue in question. This is a strong possibility, especially if parsing preferences are established while the grammar of a language is being learned, before children are reading very much. However, until large corpora of adult-to-child speech are available, there is no way to assess whether they differ from the written corpora in terms of the relative frequencies under discussion here.

Another explanation within the tuning theory framework for the non-correlation between comprehension complexity and corpus frequencies appeals to the grain size of the categories involved in tuning. Perhaps a narrower or coarser grain size will give the appropriate correlations. One possibility of a narrower grain size for which there is existing empirical motivation is a subdivision in terms of the definiteness of the NP attachment sites. Accordingly, we analyzed the middle- and high-attachment sentences from both corpora according to the definiteness of each of the three NPs. Unlike our experimental materials, naturally occurring examples are not uniformly definite or indefinite across all three sites: There are six other possible combinations of definite and indefinite sites. Furthermore, the distinction between definite and indefinite is not always easy to make. The following procedure was used in our analysis: Bare plurals, bare mass singulars, NPs with bare numerals, and NPs with indefinite articles were counted as indefinite; proper names, NPs with possessors, NPs with quantifiers such as *all* and *every*, and NPs with definite articles were counted as definite. The results of this analysis are given in Table IX.

In every subcondition except for the uniformly definite case, there are at least as many middle as high attachments. In the uniformly definite case, there are more high than middle attachments in both corpora, but these numbers are extremely small. (If referential factors were at work here, they ought to result in relatively more high than middle attachments for all cases where the highest NP is definite, but that is not the case here: There were 27 such middle attachments but only 21 high ones.) Thus, it seems that finer-grained tuning that is sensitive to definiteness still cannot explain our experimental finding of a high preference over middle in both uniformly definite and uniformly indefinite cases. The only way such a theory could work would be if (1) it turned out in an even larger sample of text that the difference between high and middle frequencies in the uniformly definite case was significant, and (2) the parser tuned its preferences for all three-

Table IX. Frequency of High and Middle Attachments According to Definitions of Three Noun Phrases (NPs)[a]

Definiteness			Brown high	Brown middle	WSJ high	WSJ middle	Total high	Total middle
NP_1	NP_2	NP_3						
def	def	def	7	3	4	3		
def	def	indef	2	5	1	1		
def	indef	def	0	2	2	3		
def	indef	indef	3	8	2	2		
Subtotal			12	18	9	9	21	27
indef	indef	indef	4	7	6	8		
indef	indef	def	1	2	7	8		
indef	def	indef	4	5	1	2		
indef	def	def	1	4	6	10		
Subtotal			10	18	20	28	30	46

[a] WSJ = *Wall Street Journal*; def = definite; indef = indefinite.

NP-site attachments based only on this one of the eight subcases. We can see no reason why tuning should operate in this way. Hence this narrower grain size does not yield a tuning explanation of our experimental complexity findings.

Of course, the fact that this particular grain size fails is not a proof that there is no similarly narrow (or narrower) grain size that makes the right predictions. In principle, there could be other narrow grain sizes that we have not identified which do allow tuning explanations of our experimental complexity findings. However, until one is identified that works, we view the tuning framework as less plausible.

There remains the possibility that the grain size is larger, collapsing over all items attaching to three-NP-site ambiguities (as in Fig. 1).[13] We examine this possibility in the following subsection.

Additional Corpus Analyses

Table X gives the frequencies of attachments to each of three preceding NP sites (matching Fig. 1) for all possible attaching categories in the Penn Treebank category tagging system for the Brown and WSJ corpora. For comparison purposes, what were listed as relative clause attachments in previous analyses are labeled S (for declarative clause), SBAR (for clause introduced by subordinating conjunction), or SBARQ in these analyses.

[13] This grain size was suggested to us by Don Mitchell.

Table X. Unfiltered Frequencies of All Possible Categories Attaching to One of Three Preceding NP Sites in the Brown and *Wall Street Journal* (WSJ) Corpora

	Attachment site					
	Brown corpus			WSJ corpus		
Attaching category[a]	NP_1	NP_2	NP_3	NP_1	NP_2	NP_3
ADJP	1	1	305	0	1	405
ADVP	1	5	37	2	5	14
NP	4	5	792	5	17	1302
PP	155	261	733	179	254	696
S	17	16	14	22	25	42
SBAR	6	2	150	29	7	162
SBARQ	0	0	1	1	1	28
SINV	0	0	0	0	0	0
SQ	0	0	0	0	0	1
VP	1	0	144	3	1	212
WHADVP	0	0	0	0	0	0
WHNP	0	0	0	0	0	0
WHPP	0	0	0	0	0	1
X	0	0	3	0	0	2
CC	54	107	357	68	88	467
Totals	239	397	2536	309	399	2865

[a] Category key: ADJP = adjective phrase; ADVP = adverb phrase; NP = noun phrase; PP = prepositional phrase; S = declarative clause; SBAR = clause introduced by subordinating conjunction; SINV = declarative sentence with subject–aux inversion; SQ = subconstituent of SBARQ (SBARQ = direct question introduced by a *wh*-phrase) excluding *wh*-word or *wh*-phrase; VP = verb phrase; WHADVP = *wh*-adverb phrase; WHNP = *wh*-noun phrase; WHNP; *wh*-noun phrase; WHPP = *wh*-prepositional phrase; X = constituent of unknown or uncertain category; CC = coordinating conjunction.

First, it should be noted that the numbers of PP and RC matches in these tables are larger than those in the tables from the studies reviewed above. As noted earlier, this is because the search procedures employed in those studies were based on heuristics which did not catch every instance of a relevant structure.

Summing over all categories, it is apparent that middle attachments are more frequent than high attachments in both corpora [Brown: $\chi^2(1) = 38.75$, $p < .001$; WSJ; $\chi^2(1) = 11.19$, $p < .001$]. So this grain size is not helpful for the tuning hypothesis. However, these frequencies come before filtering possibly inappropriate instances. Perhaps if we filter items as described earlier in this section and in the previous section, there may be a difference between middle and high attachments in the appropriate direction.

Note that almost all of the evidence for the high/middle contrast comes from the attachments of the categories PP, S, SBAR, and conjunctions (CC). To test the filtering hypothesis, it is therefore sufficient to analyze this set of categories. The frequencies for the filtered instances of these categories for the two corpora are given in Table XI.

Note that the numbers for the filtered PPs, Ss and SBARs do not match up well with the numbers reported in Gibson and Pearlmutter (1994) and Gibson and Loomis (1994), especially for PP attachment. The earlier studies found that filtered high attachments were more frequent than filtered middle attachments, whereas our analyses of *all* the matching PP attachments indicated that the reverse is actually true in the Brown corpus [$\chi^2(1) = 4.88$, $p < .05$], and that there is no difference in the filtered high and middle PP attachment frequencies in the WSJ corpus. Thus, contrary to the preliminary evidence in the second section above, frequency does not correlate with comprehension complexity for three-NP-site PP attachments either. We discuss this further in the fourth section, below.

After filtering and summing over all four categories, there are still more middle attachments overall than high attachments in both corpora, significantly so in the Brown corpus [$\chi^2(1) = 5.21$, $p < .05$], nonsignificantly so in the WSJ corpus. Hence tuning over this grain size of category predicts that people should either prefer the middle attachment site or have no strong preference. The strong preference for the high attachment in both conjunction and RC attachments (for which there are experimental data) is left unexplained. This grain size is therefore not consistent with the behavioral data.

We have explored a number of grain sizes for tuning in three-NP-site ambiguities. It was argued in the introductory section that the coarse grain

Table XI. Filtered Frequencies of the Categories PP, S, SBAR, and CC Attaching to One of three Preceding NP Sites in the Brown and *Wall Street Journal* (WSJ) Corpora[a]

	Attachment site			
	Brown corpus		WSJ corpus	
Attaching category	NP_1	NP_2	NP_1	NP_2
PP	29	47	36	36
S	5	1	4	2
SBAR	0	0	4	3
CC (conjunction)	22	36	29	37
Totals	56	84	73	78

[a] PP = prepositional phrase; S = declarative clause; SBAR = clause introduced by subordinating conjunction; CC = coordinating conjunction; NP = noun phrase.

size of tuning over two sites was not possible because of behavioral differences between two- and three-site cases. Tuning on three-NP-site cases collapsed over all attaching items was shown to be incompatible with the comprehension data, as was tuning looking only at conjoined NPs. Finer-grained tuning based on the definiteness of the three NPs also failed to match the comprehension data. Thus there seems to be a real difference between the frequency and comprehension complexity of the conjoined-NP attachment examples. It is difficult to see how the tuning hypothesis can account for this noncorrelation.

A POSSIBLE EXPLANATION OF THE FREQUENCY DATA BASED ON THE PRODUCTION OF WRITTEN TEXT

We have shown that the tuning theory does not account for the comprehension complexity results for three-NP-site conjunction and PP ambiguities, because processing results do not pattern with their input frequencies, and we have accounted for the processing findings in terms of the recency/predicate proximity theory of Gibson et al. (in press). However, we have not yet provided an explanation for the frequencies of these construction types. It cannot be claimed that complexity dictates frequency for these attachments, because the two do not correlate. The account that we propose is as follows.

We hypothesize (following Gibson & Pearlmutter, 1994) that the overall production and editing process takes account of comprehension principles in written sentence production. There are a number of ways that this might happen, which we cannot differentiate via corpus data of the kind we have analyzed. Perhaps the most obvious possibility is that written text is edited in response to comprehension difficulty that occurs when the author or an editor (re)reads the text. Another possibility is that the sentence production process engages in self-monitoring, revising complex structures before they are actually written.[14] Either of these possibilities would result in resolution frequencies patterning like comprehension complexities. Thus, this strategy accounts for the correlation between frequencies and comprehension complexity in various ambiguities reviewed in the Introduction, and for the predominance of low attachments in three-NP-site attachments across all modifier categories in our corpus counts.

[14] These processes may also apply in oral sentence production, but we do not have relevant evidence yet to test this possibility. To test this and related hypotheses, it would be useful to analyze naturally occurring spoken corpora, and compare the results of their analysis to the results presented here.

To explain the lack of correlation between the frequency and complexity of the high and middle attachments in three-NP-site conjunction and PP attachments, we assume the existence of a heuristic that prefers to put longer or heavier constituents later in the sentence than shorter ones, when the grammar allows a choice (Bever, 1970; Hawkins, 1994; Ross, 1968). In Bever's words, "Save the hardest for last." This heuristic is motivated by a desire to minimize the production and/or comprehension complexity of a structure. According to Bever, "sequences with constituents in which each subconstituent contributes information to the internal structure of the constituent are complex in proportion to the complexity of an intervening subsequence" (Bever, 1970, p. 330). In terms of production, we assume that the sentence production algorithm is top-down and that there is some memory cost associated with maintaining a category whose subconstituents have not yet surfaced. By putting the hardest/longest item later, the immediate constituent structure of high-level constituents is produced as early in the sentence as possible, thereby minimizing this cost. In terms of comprehension, a recency or locality of attachment constraint such as (3) favors local attachments. As discussed earlier, this short-term memory constraint is associated with a decay function, according to which less local sites incur greater processing load over time. By shifting heavier items later, the attachments over a whole structure are more local on average, thereby minimizing the processing load across the entire sentence, relative to an unshifted version of the same sentence.

As a result of a heuristic like this, English and other head-initial languages prefer word orders in which long or complex constituents appear as late in the sentence as possible, as shown by familiar cases of heavy-NP shift such as (10) (see Hawkins, 1994, for corpus data from several languages):

(10) a. I gave to my mother the beautiful green pendant that's been in the jewelry store window for weeks.
b. ? I gave the beautiful green pendant that's been in the jewelry store window for weeks to my mother.

In this case, (10b) is more complex than (10a) because the VP dominating the verb *give* and its arguments must be retained in memory much longer in (10b) than in (10a) because the initial subconstituent of the VP in (10b) (the heavy NP *the beautiful green pendant that's been in the jewelry store window for weeks*) is so much longer than the initial subconstituent of the VP in (10a) (the PP *to my mother*).

It turns out that this heuristic results in a greater number of middle than high attachments in corpora of production data.[15] Consider first conjunctions. The crucial point is that, among potential high- and middle-conjoined NPs, high attachment implies a first conjunct containing three NPs and two prepositions, and middle conjunction implies a first conjunct containing two NPs and one preposition, so that the first conjunct is larger for high attachment than for middle attachment. Now consider PPs. High attachment implies a first modifier containing two NPs and two prepositions, while middle attachment implies a first modifier (of NP2) containing one NP and one preposition. Thus, if the distribution of sizes of the attaching element is roughly the same across different sizes of the NP to which it is attaching, then the NP that is being attached to will be larger than the attaching constituent more often for high attachments than for middle attachments.

Given a production heuristic that encourages longer material to be later, potential structures containing a long constituent preceding a short constituent will be produced in the opposite order a percentage of the time. Following Hawkins (1994), we assume that the reordering probability is a function of length difference (or ratio) of the two constituents such that a larger difference (ratio) results in a higher probability that the larger constituent appears second. Hawkins provided corpus data to show that such a reordering function exists with respect to several constructions in English and other head-initial languages, where the grammar allows an ordering choice. In these cases, the frequency with which the item containing more words is ordered later increases as the length discrepancy between the constituents increases. Assuming such a difference-sensitive function, reordering will happen more often with high attachments than with middle attachments in the three-NP-site conjunction and PP examples because the length of the first daughter constituent in high attachments is larger than the length of the first daughter constituent in middle attachments.

For example, consider the NP conjunction in (11) and its tree structure in Fig. 3:

(11) John went to see [[the band with the drummer with the strange haircut] and [the piano soloist with the nose ring]].

For the high-attachment case in (11), the difference in word lengths of the two conjuncts is two words. Thus, given a heuristic that encourages longer material to be later, there will be some pressure to reorder (11) as (12) (see Fig. 4). Note that (12) no longer contains three potential attachment

[15] It is worth noting that the examples used in our questionnaire always involved two conjuncts of nearly equal numbers of words, so that even if this "longer–later" preference plays a role in comprehension, it would have had no effect on those results.

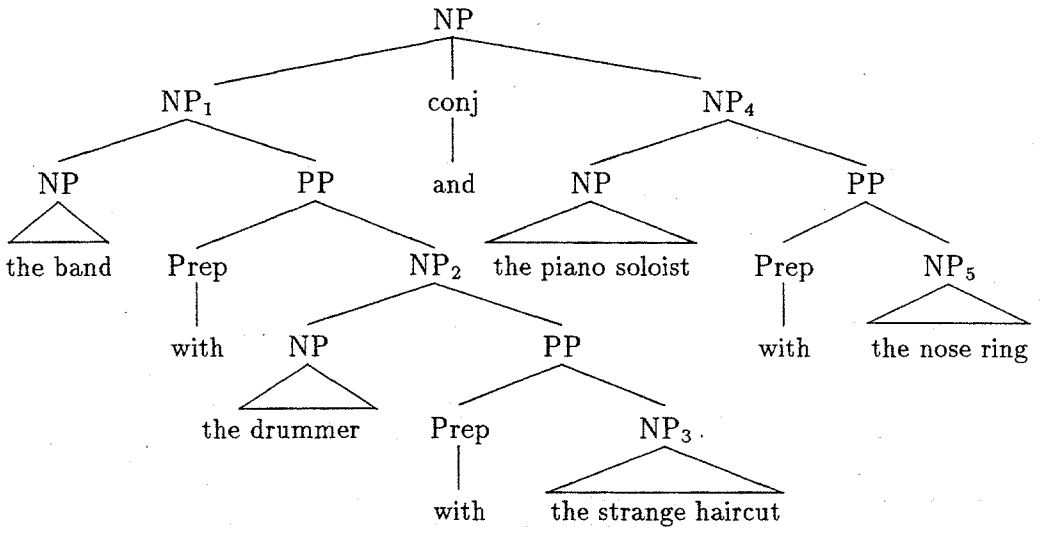

Fig. 3. High attachment of a noun phrase (NP) with an internal prepositional phrase (PP). Prep = preposition; conj = conjunction.

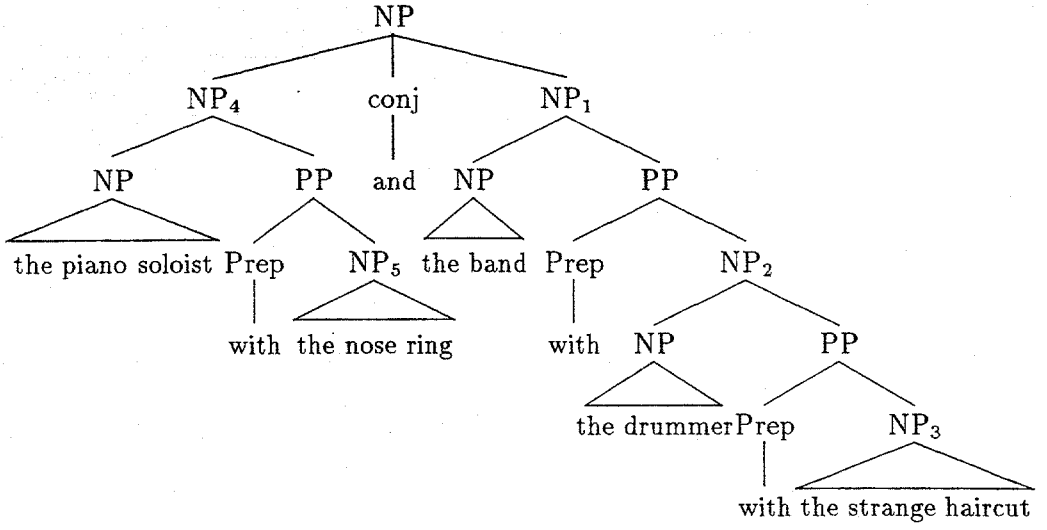

Fig. 4. After reordering the high attachment of a noun phrase (NP) with internal prepositional phrase (PP). Prep = preposition; conj = conjunction.

sites before the conjunction, and therefore would not be counted in the frequencies we computed above in the section titled Corpus Analysis.

(12) John went to see [[the piano soloist with the nose ring] and [the band with the drummer with the strange haircut]].

There should be even greater pressure to reorder the conjuncts of an example like (13), whose length difference is six words.

(13) John went to see [[the band with the drummer with the strange haircut] and [the piano soloist]].

On the other hand, there is no difference in word lengths for the two conjuncts in the middle-attachment case in (14), whose structure is shown in Fig. 5. Thus there should be no such pressure to reorder the conjoined NPs in (14).

(14) John went to see the band with [[the drummer with the strange haircut] and [the xylophonist with the long beard]].

Similarly, PPs within NPs can generally be produced in either order. (There is some tendency for argument PPs to precede adjunct PPs, but this seems to be fairly easily overridden, for example, in an NP like *a review in a journal of a new book by Chomsky*, particularly when the second PP is heavier). Thus, the proposed heuristic will also operate on PPs and will reorder high-attached PP structures more often than middle-attached ones. The result of reordering a high-attached structure will often be a structure that no longer contains a three-site attachment ambiguity, as in the following examples from the Brown corpus, where the second PP is heavier and itself contains an internal PP. In the opposite order, these would have been instances of three-NP-site high attachments.

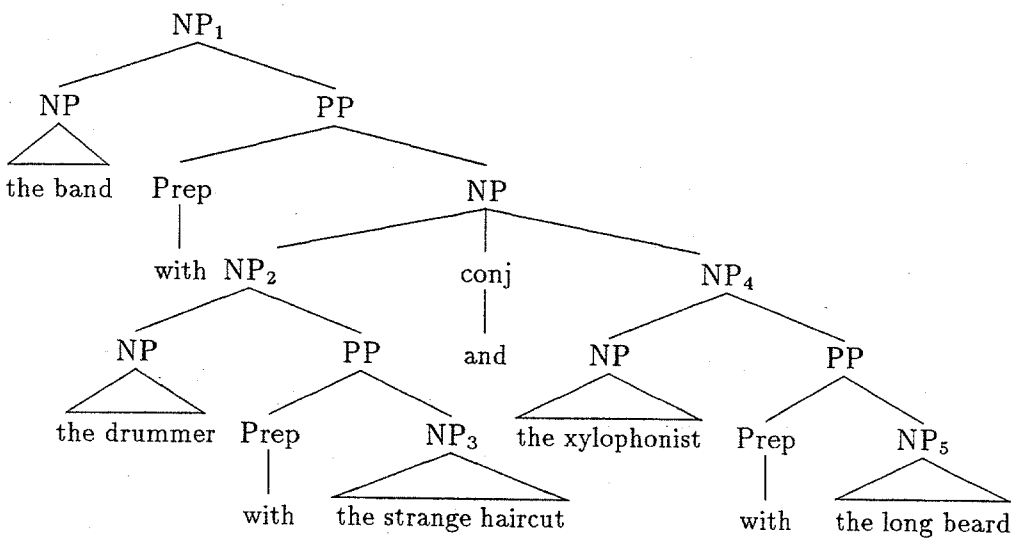

Fig. 5. Middle attachment of a noun phrase (NP) with an internal prepositional phrase (PP). Prep = preposition; conj = conjunction.

(15) a. establishment [in the future] [of a school [for the mentally retarded]]
b. divestiture [by Du Pont] [of its 63,000,000 shares [of General Motors stock]]

In general, high attachments are the most likely to be reordered given the above assumptions, which means that more potential cases of high attachment are removed from the corpus than potential cases of middle attachment, resulting in relatively more middle- than high-attachment structures in the corpus frequencies.[16] Notice that the strategies we assume to be operative in production always have the effect of reducing the comprehension complexity of the resulting sentence. However, due to the overall distribution of structures, the result of applying the strategy of shifting heavy elements is to increase the relative proportion of difficult middle attachments among the attachments to three-NP-site structures in the corpus. There is no real paradox in the fact that complexity-reducing heuristics result in a frequency distribution of structures wherein more frequent structures are sometimes more complex than less frequent ones. This can happen because certain structures, in this case middle attachments, cannot be reduced in complexity by application of the heuristic: Reordering a middle attachment results in another middle attachment, whereas reordering a high attachment can result in a less complex two-site ambiguity. Thus, frequency does not mirror complexity because the possible structures from which the production mechanism can choose, given an intended meaning, sometimes include both less and more complex options but other times do not.

In order to get independent evidence for this proposal, we conducted two additional analyses of the corpora, targeting conjoined NP structures. The first test looked at all conjunctions of two noun phrases in the two corpora under consideration, and compared the number of words in each conjunct. On the average, the second conjunct was significantly longer than

[16] We have not discussed RC attachments in this section, because the corpus frequencies are too small for us to draw any strong conclusions from them. However, we predict that there will not be more middle- than high-attaching RCs, because our proposed heuristics will not have the same effect on RCs as on PPs and conjunctions. First, an RC is substantially longer on average than a PP, so by the heaviness heuristic it will almost always appear as the later modifier, regardless of its attachment site—high attachments will not be *differentially* reordered, because all attachments will be reordered with the RC second virtually all the time. Second, predicate proximity disfavors a structure where an RC is followed by another modifier attached to the same NP, because the verb of the RC will be the closest predicate when the second modifier is attached. Thus, by both the minimization of comprehension complexity and the heaviness heuristic, we expect RCs to attach later than other modifiers and therefore to show the attachment frequencies predicted by recency and predicate proximity, namely, lows most frequent, then highs, then middles.

the first, as predicted by the production heuristic we propose. An analysis of the 5,078 conjoined NPs in the parsed Brown corpus revealed that the mean length of the first conjunct was 2.45 words, while the mean length of the second conjunct was 3.29 words [$t(10, 154) = 211.79, p << .001$]. The same analysis of the 5,474 conjoined NPs in the parsed WSJ corpus revealed mean lengths of 3.10 and 4.06 words for the first and second conjuncts, respectively [$t(10, 946) = 239.75, p << .001$].

The second test looked more specifically at the claim that is crucial for our account to explain the middle- versus high-conjunction frequencies, namely, that a greater length discrepancy between two conjuncts increases the likelihood that the longer one will occur second. To assess this, the conjoined NPs from the previous search were grouped according to the difference in length of their conjuncts in words. In each group, we calculated the percentage of sentences in which the longer NP was second. We then plotted the difference in lengths of the conjoined NPs against the percentage of instances in which the longer NP came second (shown in Figs. 6 and 7). We then calculated the correlations between the length difference and the arcsine transformation of the percentage of instances in which the longer NP came second. The linear correlation was significant for both corpora (Brown corpus: $r = .943, p << .001$; WSJ corpus: $r = .951, p << .001$).

Thus, a production heuristic like the one we propose appears to be operative, and is differentially sensitive to length, providing an account of the finding that middle conjunction and PP attachments, while more complex

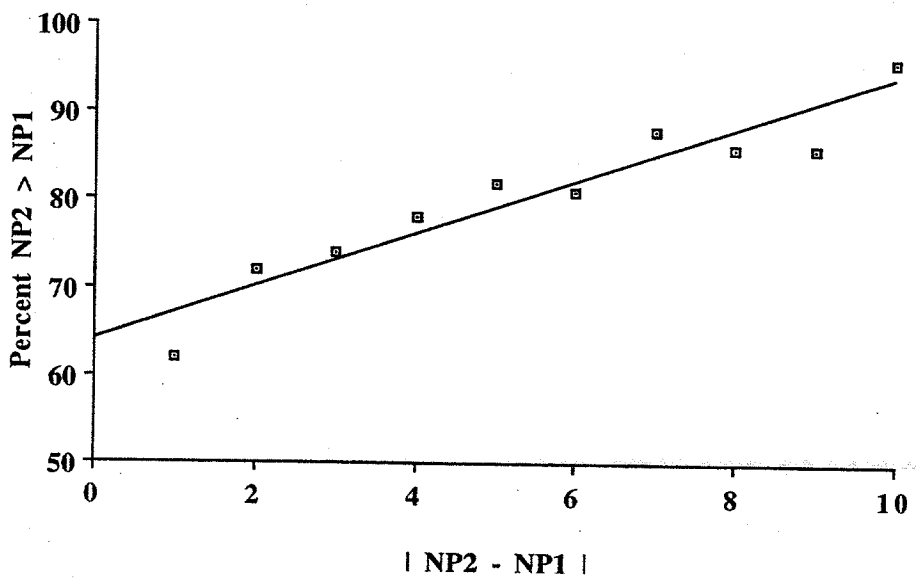

Fig. 6. The difference in lengths of conjoined noun phrases (NPs) plotted against the percentage of instances in which the longer NP comes second in the Brown corpus.

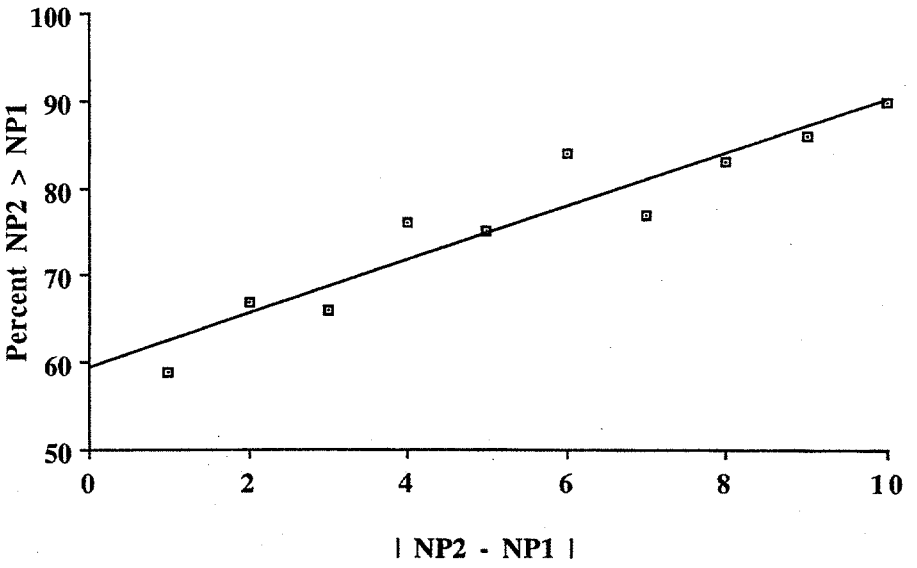

Fig. 7. The difference in lengths of conjoined noun phrases (NPs) plotted against the percentage of instances in which the longer NP comes second in the *Wall Street Journal* corpus.

to process than high attachments, are nonetheless more frequent in these corpora.

CONCLUSIONS

This paper has evaluated the tuning framework with respect to conjunctions of noun phrases in constructions with three available NP sites in English. Although the off-line comprehension experiment demonstrated a preference for high-site attachments over middle-site attachments, no frequencies in the corpora reflect this complexity ordering for any of the tuning grain sizes that we evaluated. If on-line experiments confirm this preference, then the tuning framework will be difficult to reconcile with these data. On the other hand, the recency/predicate proximity theory correctly predicts the comprehension preference data for three-NP-site cases. This theory by itself does not explain the frequency data, but these are explained by a production heuristic that, in an attempt to generally minimize complexity, shifts heavy items to the right. This heuristic results in there being relatively fewer high-conjunction and PP attachments than middle attachments.

REFERENCES

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

Anderson, J. R. (1980). *Cognitive psychology and its implications*. New York: W. H. Freeman.

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

Bever, Thomas G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Carreiras, M., & Clifton, C., Jr. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech, 36*, 353–372.

Clifton, C., Jr. (1988, August). *Restrictions on late closure: Appearance and reality*. Paper presented at the 6th Australian Language and Speech Conference, University of New South Wales, Sydney, Australia.

Clifton, C. Jr., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language, 30*, 251–271.

Corley, M., & Corley, S. (1995). *Cross-linguistic and inter-linguistic evidence for the use of statistics in human sentence processing*. Unpublished manuscript, University of Exeter, Exeter, England, and University of Edinburgh, Edinburgh, Scotland.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language processing: Psychological, computational and theoretical perspectives* (pp. 320–358). Cambridge, UK: Cambridge University Press.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition, 30*, 73–105.

Cuetos, F., Mitchell, D. C., & Corley, M. M. B. (in press). Parsing in different languages. In M. Carreiras, J. Garcia-Albea, & N. Sabastian-Galles (Eds.), *Language processing in Spanish*. Hillsdale, NJ: Erlbaum.

De Vincenzi, M., & Job, R. (1993). Some observations on the universality of the late closure strategy. *Journal of Psycholinguistic Research, 22*, 189–206.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII*. Hillsdale, NJ: Erlbaum.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology 14*, 178–210.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*, Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Gibson, E., & Loomis, J. (1994). A corpus analysis of recency preference and predicate proximity. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. (pp. 357–362) Atlanta, GA. Hillsdale, NJ: Erlbaum.

Gibson, E., & Pearlmutter, N. (1994) A corpus-based analysis of psycholinguistic constraints on prepositional phrase attachment. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 181–198). Hillsdale, NJ: Erlbaum.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (in press). Recency preference in the human sentence processing mechanism. *Cognition*.

Gilboy, E., Sopena, J. M., Clifton, C., Jr., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition, 54*, 131–167.

Hawkins, J. A. (1994). *A performance theory of order and constituency.* Cambridge, UK: Cambridge University Press.

Hays, W. (1988). *Statistics* (4th ed.). Orlando, FL: Holt, Rinehart and Winston.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics, 9,* 103–120.

Hobbs, J. R., & Bear, J. (1990). Two principles of parse preference. In *Proceedings of the Thirteenth International Conference on Computational Linguistics* (vol. 3, pp. 162–167). University of Helsinki.

Juliano, C., & Tanenhaus, M. K. (1994). A constraint-based lexical account of the subject/object attachment preference. *Journal of Psycholinguistic Research, 23,* 459–472.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition 2,* 15–47.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present day American English.* Providence, RI: Brown University Press.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32,* 692–715.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101,* 676–703.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19,* 313–330.

Merlo, P. (1994). A corpus-based analysis of verb-continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research, 23,* 435–458.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407.

Mitchell, D. C. (1994). Sentence parsing. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 375–409). New York: Academic Press.

Mitchell, D. C., & Cuetos, F. (1991). The origins of parsing strategies. *Conference proceedings: Current issues in natural language processing.* (pp. 1–12). Austin: University of Texas at Austin, CT.

Mitchell, D. C., Cuetos, F., & Corley, M. M. B. (1992, March). *Statistical versus linguistic determinants of parsing bias: Cross-linguistic evidence.* Paper presented at the 5th annual CUNY Conference on Sentence Processing, New York.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). *Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records.* Unpublished manuscript, University of Exeter, Exeter, England.

Rayner, K., Carlson, M., & Frazier, L. (1983) *The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. Journal of Verbal Learning and Verbal Behavior, 22,* 358–374.

Ross, J. R. (1968). *Universal constraints on variables.* Unpublished doctoral dissertation, MIT.

Spivey-Knowlton, M., & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55,* 227–267.

Taraban, R., & McClelland, J. R. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language, 27,* 597–632.