Semantic Integration and Syntactic Planning in Language Production

Eric S. Solomon and Neal J. Pearlmutter
Northeastern University

Send correspondence to either author:
Psychology Dept., 125 NI
Northeastern University
Boston, MA 02115

E-mail: esolomon@neu.edu, pearlmutter@neu.edu
Phone: (617) 373-3077, (617) 373-3040
Fax: (617) 373-8714

Running head: Semantic Integration
Abstract

Five experiments, using a subject-verb agreement error elicitation procedure, investigated syntactic planning processes in production. The experiments examined the influence of semantic integration — the degree to which phrases are tightly linked at the conceptual level — and contrasted two accounts of planning: serial stack-based systems and parallel activation-based systems. Serial stack-based systems rely on memory-shifting processes to coordinate ongoing planning. Memory-shifting should be easier for more integrated phrases, resulting in fewer errors. Parallel, activation-based systems, on the other hand, maintain multiple representations simultaneously in memory. More integrated phrases will be more likely to be processed together, resulting in increased interference and more errors. Participants completed stimuli like The drawing of/with the flower(s), which varied local noun number (flower(s)) and the relationship between the head (drawing) and local noun. In some constructions, the nouns were tightly integrated (e.g., of), whereas in others the relationship was looser (e.g., with, specifying accompaniment). In addition to the well-established local noun mismatch effect (more errors for plural than for singular local nouns), all experiments revealed larger mismatch error effects following tightly integrated stimuli. These results are compatible with parallel activation-based accounts and cannot be explained by serial, memory-shift-based accounts. The experiments and three meta-analyses also ruled out alternative accounts based on plausibility, argumenthood, conceptual number, clause packaging, or hierarchical feature-passing, reinforcing the general finding that error rates increase with degree of semantic integration.

Keywords: sentence production, syntactic planning, semantic integration, number agreement, syntactic features, distributivity
Talking is a seemingly effortless activity. Speakers can typically produce smooth, coherent utterances even in the absence of much conscious consideration. But this apparent ease is belied by speech error data patterns and attempts to implement computational language production systems (e.g., Bock, Eberhard, & Cutting, 2002; Dell, 1986; Fromkin, 1971; Garrett, 1975; Kempen & Hoenkamp, 1987; Vigliocco & Hartsuiker, 2002); the process of converting a thought into speech is not trivial. Models of language production reflect the complexity of this process, although they agree to a large extent on their broad organization (e.g., Bock, 1995; Bock & Levelt, 1994; Garrett, 1988). Bock and Levelt's (1994) model, for example, is separated into three components: the message level, which represents the speaker's intended meaning; grammatical encoding, which translates the intended meaning into a sequence of words expressing that meaning; and phonological encoding, which translates the sequence of words into a sequence of sounds to be uttered. The current work focuses on grammatical encoding, and in particular on syntactic planning, which takes a set of words and assigned grammatical functions (identified from the message) and creates a syntactic structure encoding hierarchical syntactic relationships, word order, and inflection.

The core problem for syntactic planning arises because of the inherent incompatibility between conceptual messages, which contain numerous co-temporaneous and often unordered elements; and word sequences, which are necessarily linear, unidimensional, and spread over time (e.g., Bock, 1987a; Levelt, 1989). Determining the order of elements is thus a critical part of the translation problem, and as a result, a variety of studies have focused on aspects of planning related to constituent ordering. This work has shown, for example, that semantically primed words and those which are more frequent, concrete, or prototypical tend to be produced earlier (Bock, 1986, 1987a, 1987b; Bock & Warren, 1985; Kelly, Bock, & Keil, 1986). More recently, examinations of phrase ordering have suggested that more prominent grammatical functions or thematic roles (e.g., agent, experiencer) are preferentially assigned to phrases in subject position (F. Ferreira, 1994), and that shorter and less complex phrases tend to be produced earlier, when possible (e.g., Arnold, Wasow, Losango, & Ginstrom, 2000; F. Ferreira & Henderson, 1998; Gibson, Schütze, & Salomon, 1996; Hawkins, 1994; Stallings, MacDonald, & O'Sheaghdha, 1998). Having multiple alternative phrase orderings available can itself also lead to faster and less error-prone utterances (V. Ferreira, 1996).

While this work has provided a range of important information about properties related to
constituent ordering and properties to which syntactic planning is sensitive, it leaves open many
questions about the mechanism which implements the planning process. In particular, the fact that
elements of the co-temporaneous message must end up spread over time and across the sequence
of words to be produced requires a system which not only applies syntactic knowledge to construct
grammatically-ordered strings, but which also is able to track multiple partial representations in
memory simultaneously (see, e.g., Kempen & Hoenkamp, 1987, for some related discussion).

One approach to this problem of coordinating multiple phrases in memory is a serial, memory-
shift-based system, in which at any given moment only a single phrase is being processed. This
approach is the analogue in sentence production to a traditional symbolic, computational parser
in comprehension (e.g., Aho & Ullman, 1972; Early, 1970/1986). The task of coordinating a
set of phrases to be uttered in such a system consists of memory-shifting operations, which shift
the current focus of processing from one phrase to another. Thus, for example, in planning the
phrase the drawing of the flowers, such a system would take information specified from the message,
indicating that the phrase should describe an object with a particular property, and would construct
an appropriate instantiating syntactic structure (e.g., a noun phrase (NP) including a modifying
prepositional phrase (PP)) using the grammatical constraints of English, inserting words from the
lexicon selected on the basis of the meaning specified in the message. In a memory-shift-based
system in particular, the process would involve first constructing the NP containing the determiner
the and the noun drawing. This phrase would then be shifted out of the current focus so that
construction of the modifier (the PP) could take place. Once the PP was built (by linking of, the,
and flowers in the appropriate syntactic structure), it would have to be attached into the original
NP, which would require additional memory-shifting. Phrases shifted out of the current focus in
such a system may still be retained if necessary for later syntactic planning processes (e.g., in a
stack-based memory), as in the case of the initial NP constructed for the drawing; or their internal
details might be dropped if they are no longer needed, as might be the case for the complete phrase
the drawing of the flowers once it was fully planned. (Whether the internal details of the syntactic
structure need to be maintained for phonological encoding or can be dropped once they are no
longer needed for syntactic planning is an open question.) But regardless of whether or not the
internal structural detail is maintained, beyond whatever resources are needed for processing the
current phrase, the system must expend resources to shift the current focus from one phrase to
another, and thus properties which make the shifting operations more difficult will increase the chance of processing errors, particularly those related to the maintenance or recall of information in memory.

An alternative mechanism for coordinating phrase-planning is a parallel, activation-based system, in which multiple partial syntactic representations are maintained in memory for simultaneous processing. The difficulty of the coordination task in such a system arises in keeping multiple phrases simultaneously active while preventing them from interfering with each other. Like a memory-shift-based approach, such a system expends resources on processing the active phrase(s). However, the system does not rely on explicit memory-shifting of phrases; instead, additional resources will be needed to prevent interference. In the case of a phrase like the drawing of the flowers, a parallel activation-based system, like a memory-shift-based one, will construct an NP PP structure, linking words to appropriate syntactic positions, using information specified from the message. Instead of shifting back and forth between construction of the NP and the PP, however, a parallel system will allow simultaneous construction of the two phrases (although the NP will likely be initiated somewhat earlier, because it will appear earlier in the utterance). Thus both the syntactic structures and the words will overlap in the timing of their activation, and to the degree that this occurs, the chance of interference between structures or between words will increase, with a corresponding increase in the chance of processing errors related to maintenance or recall of information.

These two approaches involve different tradeoffs among the requirements of translating a complex message into a string of words encoding a hierarchical structure. On the one hand, a syntactic planning system must ensure that the appropriate elements (words) are positioned and thus generated in the appropriate sequence, and on the other hand, it must maintain the ability to track connections between elements over potentially long distances. A serial, memory-shift-based system directly enforces the sequencing constraint, but this requires that additional resources be spent on shifting elements around in memory when it is necessary to track and compute relationships between non-adjacent elements. A parallel activation system relaxes the strong sequential constraint and potentially allows for easier coordination of non-adjacent elements by permitting multiple phrases to be simultaneously processed. But this comes at the cost of needing to expend resources to prevent interference between phrases.\(^1\)

These differences suggest that a critical factor to consider in trying to differentiate these systems
is the relationship between phrases to be planned. In particular, when two phrases are closely related at the message level, shifting between them in a serial system should be relatively easy: For example, their corresponding message-level elements will be closely linked and thus fewer adjustments will be needed to change the focus of processing in the syntax. This will result in easier access to the phrases and thus fewer errors. In a parallel activation-based system, however, when two phrases are closely related, they will be more likely to be active together and to share properties (again, for example, because their corresponding message-level elements will be closely linked), and this will cause them to be more likely to interfere with each other during grammatical encoding, resulting in increased processing difficulty and more errors.

We will refer to the relevant relatedness relationship between phrases to be planned as semantic integration. We consider semantic integration to be a property of the message or conceptual level of the production system and define it as how closely linked two parts of a message are within a discourse representation or mental model (e.g., Johnson-Laird, 1983). For example, although nouns like *ketchup* and *mustard* in (1a) may be similar in meaning, they are not semantically integrated (or are only very loosely integrated), because (1a) provides no information about how the two nouns are related to each other in this text. On the other hand, *bracelet* and *silver* are much more semantically integrated in (1b), because *made of* offers specific information about how the two nouns in this fragment are related, and the relation is a very tight one.

1. a. the ketchup or the mustard
   b. the bracelet made of silver

Degree of semantic integration can also be manipulated within an NP PP structure by varying just the preposition, as shown in (2). In (2a) *of* specifies a representational relationship between the initial noun *drawing* and the embedded noun *flowers*. Drawings necessarily represent some sort of content, which in this case is specified by the noun in an of-PP, and thus the two nouns are tightly integrated. On the other hand, the most likely interpretation of (2b) involves an accompaniment relationship between the same nouns. Unlike the *of* case, there is nothing necessary about the accompaniment specified by *with*, and accompaniment alone does not specify much about how the two nouns are related. Thus the two nouns are only loosely integrated.

2. a. the drawing of the flowers
   b. the drawing with the flowers
These differences in semantic integration will play out differently during the course of sentence production depending on whether the planning system is built around memory-shifting or parallel activation. On the memory-shift hypothesis, the more integrated case (2a) should create less difficulty, because shifting from processing the initial NP to the PP (and back) will be easier. On the parallel activation hypothesis, the more integrated case should be more difficult, because the NPs headed by *drawing* and *flowers* (as well as the PP) will be more likely to be active simultaneously and thus will be more likely to interfere with each other. If phrases like those in (2) are in subject position, these difficulty predictions should be measurable in terms of subject-verb agreement errors. Under the memory-shift hypothesis, because shifting will be easier in (2a) than in (2b), it will be easier to track properties of the head noun *drawing* (such as its number) in the former as well, and they will be less likely to be confused with properties of the embedded noun (*flowers*). As a result, the production of correct subject-verb agreement should be more reliable in (2a). In a parallel activation system, however, because *drawing* and *flowers* will be more likely to be active simultaneously in (2a), there will be a greater chance that the properties of *flowers* (including number) will interfere with those of *drawing*, resulting in incorrect subject-verb agreement more often.

The mechanisms of phrase planning have not previously been investigated at this level, nor have the effects of semantic integration on planning or on the implementation of subject-verb agreement. However, recent work in sentence production has examined other aspects of subject-verb agreement processes in some detail (Barker, Nicol, & Garrett, 2001; Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock et al., 2002; Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Bock & Miller, 1991; Bock, Nicol, & Cutting, 1999; Eberhard, 1997, 1999; Franck, Vigliocco, & Nicol, 2002; Hartsuiker, Antón-Méndez, & van Zee, 2001; Haskell & MacDonald, 2001; Humphreys & Bock, in press; Nicol, 1995; Thornton & MacDonald, 2003; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Butterworth, & Semenza, 1995; Vigliocco & Franck, 1999, 2001; Vigliocco, Hartsuiker, Jarema, & Kolk, 1996; Vigliocco & Nicol, 1994, 1998), and some of this work has specifically made use of subject-verb agreement error patterns to examine the nature of syntactic planning units and how transitions between them can influence agreement. In particular, Bock and Cutting (1992) suggested that one major syntactic planning unit is the clause, and they proposed that each clause in an utterance is planned as an independent, encapsulated unit. In their experiments,
participants repeated auditorily-presented subject NPs like those in (3) and then completed them as full sentences. The critical stimuli contained a singular head noun (report in (3)) followed by either a PP modifier (3a, 3b) or a full clause modifier (3c, 3d) ending in either a plural or singular local noun (fires, fire).

(3)  
  a. The report of the destructive fires
  b. The report of the destructive fire
  c. The report that they controlled the fires
  d. The report that they controlled the fire

Bock and Cutting (1992) measured subject-verb agreement errors in participants’ sentence completions following plural local nouns relative to corresponding singular local noun cases. They replicated the typical local noun mismatch error effect — increased subject-verb agreement error rates following a plural local noun relative to a corresponding singular local noun control — for PP modifier constructions like (3a vs 3b) (e.g., Bock & Miller, 1991), but they found a significantly reduced or completely eliminated effect for the clausal modifier cases (3c vs 3d). They also showed that increasing the length of the modifying PP increased the mismatch effect, but increasing the length of the modifying clause correspondingly did not affect the mismatch effect. These differences between PP and clausal modifiers indicated that, relative to embedded phrases such as PPs, properties of words within embedded clauses (e.g., noun number-marking) were much less likely to influence processes external to those clauses (e.g., subject-verb agreement) and thus suggested that each clause of an utterance (even a center-embedded one) is an independent planning unit.

Although this result does not provide direct evidence about memory-shifting versus activation-based processing mechanisms, it can help to delimit the granularity of processing units in either type of system, and the experiments presented below can be seen as extending this approach to examine constituents and constituent boundary effects within individual clauses. However, Bock and Cutting’s (1992) interpretation of their data has been challenged by Vigliocco and Nicol (1994, 1998; see also Bock et al., 2002; Franck et al., 2002; cf. Nicol, 1995), who noted that the PP versus clause difference in Bock and Cutting’s experiments was confounded with a difference in the hierarchical distance within the syntactic tree between the local noun and the NP node dominating the subject’s syntactic structure. Vigliocco and Nicol (1994, 1998) proposed that subject-verb agreement is implemented by passing the relevant features (number in English) through the subject
NP’s syntactic structure to the top NP node, and the number-marking there is then used by the processing system to mark the verb phrase (VP). Errors can arise when local noun number is incorrectly passed farther up the structure than it should be, and the number of errors which must occur in order for a subject-verb agreement error to appear is greater when the hierarchical distance from the local noun to the highest NP node is greater (as in clauses relative to PPs). Vigliocco and Nicol (1994) did not examine clauses in their own stimuli, and the hierarchical account requires a few additional assumptions to explain Bock and Cutting’s interaction between length and structure (PP versus clause), but several different studies have provided evidence suggesting that hierarchical distance is at least one relevant factor in controlling subject-verb agreement error rates (Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994, 1998; see also Nicol, Forster, & Veres, 1997; Pearlmutter, 2000, for closely related evidence from comprehension). Experiments 1–5 below thus focus on manipulations of semantic integration as a way of examining planning processes, but also consider the potential contributions of both Bock & Cutting’s clause packaging factor and hierarchical feature-passing accounts to the patterns of results. The experiments and meta-analyses also examine the relationship between semantic integration and a variety of other factors which have been related to subject-verb agreement error rates in prior studies.

Experiment 1

Experiment 1 examined the role of semantic integration in phrase planning using stimuli like those in (2), containing either an of-PP (2a) or a with-PP (2b). Each of these plural local noun versions had a corresponding control in which both the head and local noun were singular (The drawing of/with the flower). Regardless of whether the head and local nouns matched in number, the of-PP specified representational content and was more semantically integrated with the head noun than the with-PP, which specified an accompaniment relationship. Relative to corresponding singular local noun controls, the plural local noun conditions in a memory-shift system should yield fewer subject-verb agreement errors for the more integrated cases (of), while a parallel activation system predicts the opposite pattern: fewer errors for the less integrated cases (with).

Clause packaging (Bock & Cutting, 1992) does not differentiate among the of-PP and with-PP
cases; none of the subject phrases contains an embedded clause, so the relationship between the head and local nouns is always intraclausal.

Hierarchical feature-passing accounts (Bock et al., 2002; Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994, 1998) can predict several possible patterns, depending on assumptions about (1) the syntactic structures of the different conditions, and (2) whether syntactic distance depends only on the number of maximal projections (NP, VP, PP, etc.), not the total number of nodes, through which the plural feature must incorrectly pass. A third factor is whether the relevant target node of feature-passing is the subject NP maximal projection or is instead the head noun itself.²

Figure 1 shows the three syntactic structures typically considered for PP attachment into NPs (e.g., Chomsky, 1981), where the node labeled N corresponds to the position of the head noun, the PP node shows the location of the of-PP or with-PP, and the NP node at the top of each structure is the subject NP maximal projection. Obviously, if of-PPs and with-PPs attach in the same way into the subject NP, hierarchical accounts predict no difference in error rates. However, on at least some phrase structure theories, a representational of-PP will be treated as an argument of the head noun and will attach as its sister, so that it will be separated from the subject NP maximal projection by an intermediate N' node, as shown in Figure 1A. Accompaniment with-PPs are not arguments and will attach immediately below the subject NP node, either as shown in Figure 1C, or by way of adjunction, as shown in Figure 1B. The other possibility is that of-PPs might attach as in Figure 1C, and if with-PPs attach as in Figure 1B, this combination would potentially predict a difference in error rates.

Insert Figure 1 About Here

For each potential combination of different of-PP and with-PP structures, Table 1 shows the sequence of nodes (the path) through which a plural feature would have to travel in order to create a subject-verb agreement error, as well as the resulting error rate prediction, given that longer paths make errors less likely. The paths and predictions depend not only on the structures, but also on whether all nodes or only maximal projections count for distance, and on whether the
The target node for feature-passing is the highest subject NP node or is instead the head noun itself. The table shows that hierarchical feature-passing models only predict an of-PP versus with-PP error rate difference under two specific sets of assumptions. First, the of-PPs should show a lower error rate than the with-PPs only if (a) the target node is the subject NP, (b) all nodes count for distance, and (c) of-PPs attach as arguments (Figure 1A). Second, the of-PPs should show a higher error rate than the with-PPs if (a) the target node is the head noun, and (b) either the with-PPs adjoin into the subject NP (Figure 1B), or only maximal projections are counted for distance. All other combinations of possibilities yield the same feature-passing distance for the of-PPs and the with-PPs.

---

Insert Table 1 About Here

---

Method

Participants. Thirty-nine Northeastern University undergraduates participated. In this experiment and in all other experiments and surveys described below, all participants were native English speakers and received either class credit or payment ($10). No participant contributed data to more than one experiment/survey.

Materials and Design. Twenty-four stimulus sets like that shown in (4) were constructed. Each sentence preamble consisted of a head NP (e.g., *The drawing* in (4)) followed by a preposition (*of* or *with*) and a local NP (*the flower(s)*). The head NP was always singular, and the four different versions of an item were created by varying the local NP number and the preposition. All head NPs were like *the drawing* in that they contained nouns that had representational content.

(4) a. The drawing of the flower
   b. The drawing of the flowers
   c. The drawing with the flower
   d. The drawing with the flowers

In addition to the experimental items, 60 filler preambles were included. Twenty of the fillers
consisted of an NP PP sequence like that for the experimental items, and 18 of these had a plural head noun. The remaining fillers had a variety of structures. The experimental and filler stimuli were combined to form 4 counterbalanced 84-item lists. Each list included the 60 fillers and exactly one version of each of the 24 experimental preambles.

**Semantic integration rating.** For all experimental items in Experiments 1–5, a separate group of 96 Northeastern University undergraduates rated the preambles for how closely linked the head and local nouns were within each phrase. The 104 items were used to create 48 counterbalanced lists. Each list contained 80 items, with 16 items per page. The order of the pages was randomized separately for each participant. Each version of each item was rated by at least 10 participants, using a scale of 1 (not linked) to 7 (tightly linked). The instructions included example phrases (*the ketchup or the mustard and the bracelet made of silver*) and indicated that although *ketchup* and *mustard* are similar in meaning, the particular example phrase does not provide any close connection between the two words, in contrast to *bracelet* and *silver* in the other example.

Mean semantic integration ratings for the Experiment 1 stimuli are shown in Table 2. These ratings were analyzed in a 2 (preposition) × 2 (local noun number) ANOVA with items as the random factor (all analyses of integration and of plausibility ratings below also used items as the random factor), which showed that, as intended, the of-PP items were rated as more tightly linked than the with-PP items ($F(1,23) = 43.86$, $MS_e = .28$, $p < .001$). There was no effect of local noun number, and there was no interaction between preposition and local noun number ($Fs < 1$).

---

**Insert Table 2 About Here**

---

**Plausibility norming.** To examine if any effects of preposition or local noun number on error rates might be due to general plausibility differences, a separate group of 80 Northeastern University undergraduates rated the preambles for plausibility. The 24 items were combined with 43 fillers with similar syntactic structures to create 4 counterbalanced 67-item lists with 16 or 17 items per page. Ratings were conducted for the four versions of each item, as shown in (4). Each list had
a different random ordering of items, and the pages of a list were randomly ordered separately for each participant. Participants rated each item using a scale of 1 (implausible) to 7 (plausible). Participants were instructed to judge plausibility and were given a clearly plausible and a clearly implausible example.

The mean plausibility ratings for Experiment 1 are shown in Table 2. These ratings were analyzed in a 2 (preposition) × 2 (local noun number) ANOVA, which showed that of-PP items were rated as more plausible than with-PP items (\(F(1,23) = 20.68, MS_e = .72, p < .001\)) and yielded an interaction between preposition and local noun number (\(F(1,23) = 12.43, MS_e = .13, p < .01\)). For the of-PP items, the singular versions were judged to be more plausible than the plural versions (\(t(23) = 2.49, p < .05\)), and for the with-PP items, the plural versions were judged to be more plausible than the singular versions, though this difference was marginal (\(t(23) = 1.97\)). There was no main effect of local noun number (\(F < 1\)).

**Procedure.** Participants were run individually. Each preamble was presented visually on a computer screen, and the participant’s task was to repeat the preamble out loud with an ending to form a complete sentence. No instructions or feedback were given about the content or structure of the completions, except that they should form complete sentences.

On each trial, a fixation cross appeared at the left edge of the screen for 1000 ms, followed by the presentation of the preamble (with its first character positioned where the fixation cross had been). Preambles were presented for the larger of 1000 ms or 40 ms per character. The participant repeated back the preamble as rapidly as possible along with a completion. After the preamble display, each trial ended with a blank screen for 3000 ms, followed by a message instructing the participant to press the space bar for the next trial. An IBM-compatible computer running the MicroExperimental Laboratory (MEL) software package (Schneider, 1988) presented the preambles, and the participants’ responses were recorded uncompressed onto CD-R for analysis, using a Shure SM58 microphone connected to a Mackie 1202-VLZ Pro mixer/preamp and a Fostex CR300 CD recorder.

In total, the experiment consisted of 84 trials preceded by 10 practice items similar in length and complexity. If, at any point, participants’ speech rates dropped noticeably during the task, the experimenter encouraged them to talk faster.
Scoring. All completions were transcribed and then assigned to one of four main scoring categories. A completion was scored as a correct response when a participant repeated the preamble correctly, said it only once, produced an inflected verb as the first word of the completion, and used a verb form that was correctly marked for number. Completions were scored as errors when all of the above criteria for correct responses were met, except that the verb form failed to agree in number with the subject. The third category, uninflected responses, contained completions that met all of the criteria for correct responses except that the verb had no overt inflection for number. Finally, responses were scored as miscellaneous when a participant made an error in repeating the preamble, when a verb did not immediately follow the preamble, or when a response did not meet the criteria for any of the other categories. Trials on which the participant made no discernible response at all were excluded from all analyses. In addition to these scoring categories, we also noted cases where speakers produced dysfluencies (e.g., pauses, coughs) during or immediately after the preamble. If a speaker produced a dysfluency and then continued to complete the sentence in a way that matched the criteria for one of the first three scoring categories above, then both the dysfluency and the scoring category were recorded. On trials scored in the miscellaneous category, dysfluencies were not counted separately.

Results

Separate analyses by condition were performed on the percentage of error responses (out of errors plus correct responses), the number of uninflected responses, and the number of miscellaneous responses. For each response type, we performed two 2 (preposition) × 2 (local noun number) analyses of variance (ANOVAs), one with participants ($F_1$) and another with items ($F_2$; Clark, 1973) as the random factor. The statistics and error rates reported in all of Experiments 1–5 are for analyses excluding dysfluency cases (except for the miscellaneous category). Unless otherwise noted, the statistical patterns were identical if dysfluency cases were included.

The percentage of agreement error responses (out of errors plus correct responses) and the numbers of each response are shown in Table 3. Speakers produced more agreement errors in the plural local noun conditions than in the singular local noun conditions ($F_1(1, 38) = 18.48, MS_e = 394.47, p < .001; F_2(1, 23) = 20.54, MS_e = 240.22, p < .001$), and more errors were produced following of-PPs than with-PPs ($F_1(1, 38) = 26.66, MS_e = 70.11, p < .001; F_2(1, 23) = 9.75, MS_e$
= 125.75, \( p < .01 \)). Most importantly, local noun number and preposition interacted \((F_1(1, 38) = 25.05, MS_e = 63.97, p < .001; F_2(1, 23) = 7.35, MS_e = 108.23, p < .05)\), because the head-local mismatch effect for the of-PP cases \((F_1(1, 38) = 24.92, MS_e = 315.57, p < .001; F_2(1, 23) = 19.31, MS_e = 251.09, p < .001)\) was larger than for the with-PP cases \((F_1(1, 38) = 7.20, MS_e = 142.87, p < .05; F_2(1, 23) = 9.07, MS_e = 97.36, p < .01)\). This same pattern was present if the number of errors (rather than the percentage of errors) was analyzed.

---

Insert Table 3 About Here

---

No significant main effects or interactions were found for the number of uninflected responses \((Fs < 1)\). There were more miscellaneous responses for with-PPs than for of-PPs \((F_1(1, 38) = 13.32, MS_e = .624, p < .01; F_2(1, 23) = 11.71, MS_e = 1.15, p < .01)\), but there was no effect of local noun number \((Fs < 1)\) and no interaction \((F_1(1, 38) = 2.85, MS_e = .441, p > .10; F_2(1, 23) = 2.94, MS_e = .694, p = .10)\).

**Discussion**

Experiment 1 revealed two primary results. First, speakers produced more errors in the plural local noun conditions than in the singular local noun conditions, replicating the head-local mismatch effect found in nearly all other studies using singular-head NP PP constructions (e.g., Bock & Miller, 1991). Second, the head-local mismatch effect was larger in the of-PP conditions than in the with-PP conditions, indicating that when the head and local nouns mismatch in number, more integrated cases yield higher error rates. This is the pattern predicted by parallel activation-based systems, in which more integrated phrases will tend to be active simultaneously and will thus be more likely to interfere with each other. In the constructions of Experiment 1, the interference manifests as subject-verb agreement errors because the number-marking of the (plural) local NP interferes with that of the (singular) head NP.

The Experiment 1 interaction pattern is not compatible with memory-shift-based systems, because they predict that shifting will be easier, and thus successful implementation of subject-verb agreement will be more likely, in more integrated cases (the of-PPs). The clause packaging expla-
nation for differences in subject-verb agreement error rates between clauses and PPs also cannot account for the results, because it fails to differentiate the of-PP and with-PP conditions at all. Hierarchical feature-passing accounts as proposed in previous studies, in which the target for feature-passing is the subject NP maximal projection, fail as well, as shown in the top half of Table 1. These accounts either predict no difference between the of-PP and with-PP mismatch error rates, or they predict that the of-PPs should yield lower rates than the with-PPs. Some versions of the alternative hierarchical feature-passing account, in which the target of feature-passing is the head noun, can account for the results, depending on several additional assumptions (see the bottom half of Table 1). We will examine these accounts in more detail in later experiments.

The Experiment 1 results, while providing support for a parallel activation-based planning system over a memory-shift-based system, may still be somewhat limited, however, because of two concerns: First, the stimuli were all like those in (4), contrasting a representational relationship in the of-PP conditions with an accompaniment relationship in the with-PP conditions, and it could be that the mismatch error interaction resulted from some more specific contrast between these two semantic relationships rather than a general contrast between tighter and looser integration. Second, the plausibility norms indicated that local noun number interacted with preposition. The interaction does not map precisely onto the interaction pattern in error rates — the singular of-PP condition was more plausible and yielded fewer errors than the corresponding plural, but the singular with-PP condition was less plausible while still yielding fewer errors than the corresponding plural — but it could certainly contribute. Experiment 2 was designed to address these concerns.

**Experiment 2**

Experiment 2 further examines the effects of semantic integration on phrase planning, using a different kind of semantic relationship for the preposition of, as in The chauffeur of the actor(s). Of in this case specifies a functional relationship between the head and local nouns, such that the chauffeur is performing some sort of function for the actor(s). This is in contrast to the representational relationship specified by of in Experiment 1. However, just as in Experiment 1, the head and local nouns are tightly integrated. In The chauffeur with the actor(s), on the other
hand, with specifies a less integrated accompaniment relationship between the nouns, also as in Experiment 1.

Memory-shift and parallel-activation systems make the same predictions for Experiment 2 as for Experiment 1: More integrated cases (the of-PPs) should yield a smaller mismatch error effect than less integrated cases (the with-PPs) in a memory-shift system; a parallel activation system predicts a larger mismatch effect in the more integrated cases. Also as in Experiment 1, clause packaging provides no basis for differentiating the conditions, and the predictions of hierarchical feature-passing accounts are as shown in Table 1, with the possibilities for the functional of-PPs replacing those for the representational of-PPs.

Method

Participants. Fifty-two Northeastern University undergraduates participated.

Materials and Design. Twenty stimulus sets like that shown in (5) were constructed. As in Experiment 1, the head NP was always singular, and the four different versions of an item were created by varying the local NP number and the preposition. All of-PPs specified the target of a functional relationship.

(5)  a. The chauffeur of the actor
    b. The chauffeur of the actors
    c. The chauffeur with the actor
    d. The chauffeur with the actors

In addition to the 20 experimental items, the 60 fillers from Experiment 1 were used again, along with 4 new filler items (none of which consisted of an NP PP sequence). The experimental and filler stimuli were combined to form 4 counterbalanced 84-item lists. Each list included all 64 fillers and exactly one version of each of the 20 experimental preambles.

Semantic integration rating. Each version of the 20 stimulus sets was included in the semantic integration survey described in Experiment 1. The mean semantic integration ratings for the Experiment 2 stimuli are shown in Table 4. An ANOVA on the ratings yielded effects of preposition \( (F(1, 19) = 6.01, MS_e = .19, p < .05) \), because the of-PP versions were rated as being more tightly
linked than the with-PP versions, and of local noun number \(F(1, 19) = 6.63, MS_e = .27, p < .05\), with singular items rated higher than plurals. There was no interaction \(F < 1\).

\[\text{Insert Table 4 About Here} \]

\textbf{Plausibility norming.} Fifty-three Northeastern University undergraduates rated the experimental preambles for plausibility. Twenty-four stimulus sets like that shown in (5) were used to create 4 counterbalanced one-page lists. All items had the same NP PP structure, and all of the of-PPs specified a functional relationship between two animate nouns. Ratings were conducted for the four versions of each item, as shown in (5). As in the plausibility norming for Experiment 1, each list had a different random ordering of items, and participants rated each item using a scale of 1 (implausible) to 7 (plausible). The 20 highest-rated items were used as the critical stimuli for the experiment, and their mean plausibility ratings are shown in Table 4. A 2 (preposition) \(\times\) 2 (local noun number) ANOVA on the ratings showed no main effects and no interaction \((Fs < 2, ps > .20)\).

\textbf{Procedure and Scoring.} The procedure and response scoring were as in Experiment 1.

\textbf{Results}

The percentage of agreement errors and the numbers of each response type are presented in Table 5. Speakers produced more agreement errors in the plural local noun conditions than in the singular local noun conditions, though this comparison was marginal by items \((F_1(1, 51) = 17.57, MS_e = 393.94, p < .001; F_2(1, 19) = 4.03, MS_e = 483.57, p < .10)\), and there was no effect of preposition \((F_1(1, 51) = 2.41, MS_e = 497.95, p > .10; F_2 < 1)\). However, as in Experiment 1, local noun number and preposition interacted \((F_1(1, 51) = 4.87, MS_e = 438.77, p < .05; F_2(1, 19) = 7.37, MS_e = 221.66, p < .05)\). There was a reliable local noun mismatch effect for the of-PPs \((F_1(1, 51) = 15.02, MS_e = 557.44, p < .001; F_2(1, 19) = 13.17, MS_e = 271.53, p < .01)\), but not for the with-PPs \((F_1(1, 51) = 2.48, MS_e = 275.26, p > .10; F_2 < 1)\). If the number of errors (rather than
the percentage of errors) was analyzed, the effect of preposition was marginal by participants and significant by items, and the effect of local noun number was significant both by participants and by items. The interaction did not change if the number of errors was analyzed.

Insert Table 5 About Here

No significant effects or interactions were present for uninflected responses ($F$s < 1) or miscellaneous responses ($F$s < 2, $p$s > .20).

Discussion

As in Experiment 1, speakers produced more errors in the plural local noun conditions than in the singular local noun conditions, and local noun number and preposition interacted. The interaction pattern matched that in Experiment 1, and thus these results are again compatible with a parallel activation system for phrase planning, but not with a memory-shift system and not with either clause packaging or hierarchical feature-passing accounts in which distance is measured to the subject NP maximal projection. The Experiment 2 conditions were rated as equally plausible, so plausibility cannot explain the error effects, and the pattern is not specific to the particular semantic roles compared in Experiment 1 (representational and accompaniment).

While the error rate patterns were very similar in Experiments 1 and 2, one difference is that the uninflected response rate was much higher in Experiment 2 than in Experiment 1. However, this is not surprising given the difference in animacy of the subject NPs in the two experiments. Bock and Miller (1991) noted the same effect and suggested that it resulted from a wider range of content verbs being available for animate than for inanimate subject NPs, and participants tending not to inflect content verbs (because they are usually used in the past tense).
Experiment 3

Despite the converging evidence from Experiments 1 and 2, there are at least two other confounds not yet ruled out: First, obviously, Experiment 1 and 2 compared the same two prepositions, *of* and *with*. Although *of* assigned (or transmitted) different semantic roles in the two experiments, the interaction might nevertheless be specific to these two prepositions. Second, argumenthood status of the local PP was probably confounded with semantic integration in Experiments 1 and 2. Determining the argument status of PPs attaching into NPs is not straightforward, and the argument/adjunct distinction may in fact be a continuum in such cases (see, e.g., Schütze, 1995; Schütze & Gibson, 1999; cf. Boland, Blodgett, & Ainsworth-Darnell, 1999; Clifton, Speer, & Abney, 1991; Frazier & Clifton, 1996), but to the extent that a difference exists, the with-PPs in both experiments are adjuncts, while the of-PPs are more likely to be treated as arguments. Thus an account in which subject-verb agreement errors occur more often when the phrase containing the local noun is an argument rather than an adjunct could explain the results of Experiments 1 and 2.\(^5\) A third possibility is that, as shown in the bottom half of Table 1, some versions of a hierarchical feature-passing account in which distance is measured to the head noun could explain the Experiment 1 and 2 results.

Experiment 3 examined these confounds by replacing *of* with a different preposition, *for*, in the Experiment 2 stimuli (e.g., *The chauffeur for the actor(s)*). While the for-PPs in these cases express the same semantic role as the of-PP versions, and maintain the same degree of semantic integration, the for-PPs are adjuncts rather than arguments on most tests (Schütze & Gibson, 1999). The with-PP versions were unchanged from Experiment 2 and were thus also adjuncts, so if argumenthood status of the local phrase were the factor responsible for the interaction in the prior results, no interaction should appear in this experiment. On the other hand, because Experiment 3 compared more integrated functional relationship cases to less integrated accompaniment relationship cases, the memory-shift and parallel activation system possibilities make the same predictions in terms of semantic integration as in Experiments 1 and 2. Clause packaging again predicts no interaction, and the same is true for all hierarchical feature-passing accounts, because both PPs will attach into the subject NP structure in the same way (either as in Figure 1B or 1C).
Method

Participants. Sixty Northeastern University undergraduates participated.

Materials and Design. The same experimental stimuli and design from Experiment 2 were used, except that the preposition of was replaced by for.

In addition to the 20 experimental items, the same 64 filler preambles from Experiment 2 were used. The experimental and filler stimuli were combined to form 4 counterbalanced 84-item lists. Each list included all 64 fillers and exactly one version of each of the 20 experimental preambles.

Semantic integration rating. The mean semantic integration ratings for Experiment 3 are shown in Table 6. They were obtained as described in Experiment 1. (The with-PP stimuli were not rated twice; the data from Experiment 2 were used again.) A 2 × 2 ANOVA on the ratings indicated that the for-PPs were more tightly linked than the with-PPs ($F(1, 19) = 18.41, MS_e = .12, p < .001$). There was also a marginal effect of local noun number ($F(1, 19) = 4.17, MS_e = .18$), because the singular versions were rated as more integrated than the plural versions. Preposition and local noun number did not interact ($F(1, 19) = 1.08, MS_e = .18, p > .30$).

=----------------------------
Insert Table 6 About Here
=----------------------------

Plausibility norming. Forty-four Northeastern University undergraduates rated the preambles for plausibility. The 20 for-PP versions of the experimental items were paired with 20 locative items (described in Experiment 4 below) and 42 fillers with similar syntactic structures, to create 4 counterbalanced 62-item lists with 14 or 16 items per page. Twenty-four of the 42 fillers were the with-PP items from Experiment 4, as described below. As in Experiment 1, each list had a different random ordering of items, and the pages of a list were randomly ordered separately for each participant. The instructions and rating scale were also as in Experiment 1.

The mean plausibility ratings are shown in Table 6. A 2 (preposition) × 2 (local noun number) ANOVA on the ratings showed that the for-PPs were more plausible than the with-PPs ($F(1, 19) =$
7.47, \( MS_e = .24, p < .05 \), but there was no main effect of local noun number nor an interaction \((F_s < 1)\).

**Procedure and Scoring.** The procedure and response scoring were as in Experiment 1.

**Results**

The percentage of agreement errors and the numbers of each response type are presented in Table 7. Speakers produced more agreement errors in the plural local noun conditions than in the singular local noun conditions \((F_1(1, 59) = 10.52, MS_e = 429.79, p < .01; F_2(1, 19) = 10.19, MS_e = 368.75, p < .01)\), and there was no effect of preposition \((F_1(1, 59) = 2.17, MS_e = 270.49, p > .10; F_2(1, 19) = 1.16, MS_e = 423.72, p > .25)\). However, as in Experiments 1 and 2, local noun number and preposition interacted, although this was marginal by items \((F_1(1, 59) = 5.36, MS_e = 257.08, p < .05; F_2(1, 19) = 3.81, MS_e = 520.65)\). Similar to Experiment 2, there was a reliable local noun effect for the for-PPs \((F_1(1, 59) = 10.56, MS_e = 515.64, p < .01; F_2(1, 19) = 13.82, MS_e = 405.20, p < .001)\), but not for the with-PPs \((F_1(1, 59) = 2.65, MS_e = 171.22, p > .10; F_2 < 1)\). If dysfluency cases were included in the analysis, the effect of local noun number was significant by participants but marginal by items, and the interaction between local noun number and preposition was significant both by participants and by items. Otherwise, the same pattern appeared if dysfluencies were included. If the number of errors (rather than the percentage of errors) was analyzed, the pattern matched that for percentage of errors, except that the local noun effect for the with-PPs was reliable by participants, though not by items.

-----------------------------

Insert Table 7 About Here

-----------------------------

No significant effects or interactions were present for uninflected responses \((F_s < 3, ps > .10)\). For miscellaneous responses, the main effect of local noun number was reliable by participants and marginal by items \((F_1(1, 59) = 6.24, MS_e = .487, p < .05; F_2(1, 19) = 3.92, MS_e = 2.32)\). There was no effect of preposition and no interaction \((F_s < 2, ps > .30)\).
Discussion

As in the first two experiments, speakers produced more errors in the plural local noun conditions than in the singular local noun conditions, and there was an interaction between local noun number and preposition, with a larger mismatch effect for more integrated cases than for less integrated cases. Because both the for-PPs and the with-PPs are adjuncts in these stimuli, an influence of argumenthood status on subject-verb agreement cannot be responsible for the effects, nor can they be due to some difference specific to of versus with or to clause packaging or hierarchical feature-passing effects. Instead, the results again support a parallel activation-based system for syntactic planning, in which more integrated cases yield greater interference between phrases and thus higher agreement error rates. Experiment 4 provided an additional test of the generality of this result by further examining issues related to different prepositions.

Experiment 4

Like Experiment 3, this experiment manipulated the semantic role of the attaching PP, comparing more and less integrated cases, and both types of PPs again were adjuncts. Unlike prior experiments, however, the same preposition (with) was used for both roles to be assigned, and the difference was created by varying the content of the local NP: In the tightly integrated conditions, the PP specified an attribute of the head noun (e.g., The pizza with the yummy topping(s)). In the loosely integrated conditions, the PP received an accompaniment interpretation (e.g., The pizza with the tasty beverage(s)). The predictions of the memory-shift and parallel activation-based systems for these stimuli are as in prior experiments, and clause packaging and all hierarchical accounts again predict no difference between conditions.

Experiment 4 also included a set of locative PP stimuli, created by replacing the preposition from the experimental stimuli used in Experiments 2 and 3 with a locative preposition (e.g., The chauffeur beside the actor(s)). Like all of the with-PP stimuli, locative PPs are adjuncts. They provide a set of loosely integrated stimuli which do not involve the preposition with, so they can help to test the generality of effects of integration.
Method

Participants. Forty-seven Northeastern University undergraduates participated.

Materials and Design. For the main experimental stimuli, twenty-four stimulus sets like that shown in (6) were constructed. The head NP was always singular and the preposition was always with; the four different versions of an item were created by varying the local NP number and the semantic role of the PP. Depending on the local NP, the PP specified either an attribute of the head noun as in (6a) and (6b), or accompaniment, as in (6c) and (6d). The local NPs were matched for number of syllables within each stimulus set.

(6) a. The pizza with the yummy topping
   b. The pizza with the yummy toppings
   c. The pizza with the tasty beverage
   d. The pizza with the tasty beverages

Included with the with-PP items were 20 preambles created by substituting a locative preposition (behind, beside, or near) in the critical items from Experiments 2 and 3. These stimuli varied local NP number, and the same locative preposition was used in both local NP versions of a given item.

Forty additional filler preambles, 18 of which contained plural head NPs, were also included. Most of these fillers were also included in Experiment 3. The attribute/accompaniment, locative, and filler stimuli were combined to form 4 counterbalanced 84-item lists. Each list included all 40 fillers and exactly one version of each of the 24 with-PP and 20 locative preambles.

Semantic integration rating. Mean semantic integration ratings for the with-PP and locative stimuli are shown in Table 8. An ANOVA on the with-PP items showed that the attribute versions were more tightly linked than the accompaniment versions \(F(1, 23) = 73.07, MS_e = 1.62, p < .001\), and there was no effect of local noun number nor an interaction \(F_s < 1\).

---

Insert Table 8 About Here

---
**Plausibility norming.** Each version of the 24 with-PP stimulus sets was included in the plausibility survey described in Experiment 3 along with the locative stimuli. The mean plausibility ratings are shown in Table 8. The with-PP items were submitted to a 2 (semantic role) × 2 (local noun number) ANOVA, which showed that attribute PPs were rated as more plausible than accompaniment PPs \((F(1, 23) = 45.65, MS_e = 1.19, p < .001)\). There was no effect of local noun number, and no interaction between local noun number and semantic role \((Fs < 2, ps > .20)\). There was no difference in plausibility for the locative-PP pairs \((t(19) = 1.03, p > .30)\).

**Procedure and Scoring.** The procedure and response scoring were as in Experiment 1.

**Results**

**Attribute/accompaniment stimuli.** The percentage of agreement errors and the numbers of each response type are presented in Table 9. Speakers produced more agreement errors in the plural local noun conditions than in the singular local noun conditions \((F_1(1, 46) = 40.06, MS_e = 478.42, p < .001; F_2(1, 23) = 22.92, MS_e = 313.96, p < .001)\), and there was no effect of semantic role \((F_1(1, 46) = 2.25, MS_e = 517.16, p > .10; F_2(1, 23) = 1.45, MS_e = 339.91, p > .20). Moreover, as in Experiments 1–3, there was a significant interaction between local noun number and semantic role \((F_1(1, 46) = 8.63, MS_e = 467.95, p < .01; F_2(1, 23) = 11.04, MS_e = 167.81, p < .01)\), because the local noun effect for the attribute PPs \((F_1(1, 46) = 35.17, MS_e = 580.14, p < .001; F_2(1, 23) = 24.12, MS_e = 338.98, p < .001)\) was larger than that for the accompaniment PPs \((F_1(1, 46) = 7.65, MS_e = 336.23, p < .01; F_2(1, 23) = 6.11, MS_e = 142.79, p < .05)). The same pattern appeared if the number of errors (rather than the percentage of errors) was analyzed.

----------------------------------
Insert Table 9 About Here
----------------------------------

Speakers produced more uninflected responses in the attribute cases than in the accompaniment cases \((F_1(1, 46) = 17.37, MS_e = 1.33, p < .001; F_2(1, 23) = 8.68, MS_e = 5.22, p < .01)\), and they produced more uninflected tokens in the plural local noun conditions than in the singular local noun conditions, though this comparison was only significant by items \((F_1(1, 46) = 2.05, MS_e = 2.03,
\[ p > .10; \] \[ F_2(1, 23) = 4.54, MS_e = 1.80, p < .05 \]. There was no significant interaction \((Fs < 1)\), nor were there any significant differences for miscellaneous responses \((all \ F_s < 1)\).

**Locative stimuli.** The percentage of agreement errors and the numbers of each response type for the locatives are presented in Table 9. The head-local mismatch error effect was marginal by participants and significant by items \((F_1(1, 46) = 3.86, MS_e = 483.83, p = .055; F_2(1, 23) = 11.22, MS_e = 129.66, p < .01)\). If the number of errors \((rather than percentage of errors)\) was analyzed, the effect was significant both by participants and by items.

There were more uninflected responses for singular than for plural local nouns \((F_1(1, 46) = 15.64, MS_e = 1.26, p < .001; F_2(1, 19) = 12.32, MS_e = 3.75, p < .01)\), and there were more miscellaneous responses for plural than for singular local nouns \((F_1(1, 46) = 15.92, MS_e = 1.24, p < .001; F_2(1, 19) = 9.02, MS_e = 5.12, p < .01)\).

**Discussion**

In Experiment 4, speakers produced larger mismatch error effects in the more integrated attribute than in the less integrated accompaniment conditions. These results provide further support for the importance of semantic integration in syntactic planning and provide additional evidence against explanations in terms of clause packaging, hierarchical feature-passing, and argumenthood. The locative stimuli, which were intermediate in integration between the attribute and accompaniment items, also fit this pattern, as the locative mismatch error effect \((10\%)\) was intermediate between the attribute \((26\%)\) and accompaniment \((7\%)\) mismatch effects.

One other notable aspect of the results is the relatively high error rate for the singular local noun accompaniment items, compared to the corresponding with-PP conditions in Experiments 1–3. This most likely reflects the possibility of a conjunctive reading for *with*, which might have inflated the probability of interpreting the subject NP as plural. However, this possibility should apply equally well in the plural local noun condition, increasing its error rate correspondingly, so that the singular local noun condition still serves as an appropriate control for the corresponding plural case.

Although this conjunctive interpretation may also have been available for the with-PPs in Experiments 1–3, it is probably less likely in those experiments: In Experiment 4, the head and
local nouns in the accompaniment conditions were chosen to be similar and to be objects that would be encountered in the same situation (e.g., pizza and beverage). In Experiment 1, on the other hand, the head and local nouns were usually quite different, because the head noun was always representational, whereas the local noun typically was not. Similarly, in Experiments 2 and 3, while the head and local nouns were both animate, they were again relatively dissimilar (e.g., chauffeur and actor). Thus the conjunctive reading was probably most likely to appear in Experiment 4.

Experiment 5

While clause packaging and hierarchical feature-passing cannot account for the range of results of Experiments 1–4, these experiments also cannot rule out the possibility that one or the other does have an effect on subject-verb agreement processes. None of Experiments 1–4 included a manipulation in which effects of clause packaging should have appeared, and the same is true for at least some versions of hierarchical feature-passing (see Table 1), or all versions if all of the Experiment 1–4 PPs attach into the subject NP in the same way. Because there is support in prior studies for clause packaging (Bock & Cutting, 1992) and for hierarchical feature-passing accounts (Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994), an important question is whether these might be needed in addition to semantic integration, in order to explain subject-verb agreement results.

To examine this question, Experiment 5 contrasted PP and embedded clause constructions like those used by Bock and Cutting (1992), in which both clause packaging and hierarchical feature-passing accounts predict a difference: smaller head-local mismatch effects for clauses than for PPs. As in Experiments 1–4, semantic integration was also manipulated, in order to examine how it would interact with the other factors. The stimuli included both relative clause (RC) and sentence complement (SC) constructions, which also permitted an additional test of the argumenthood account: Experiments 3 and 4 controlled argumenthood while manipulating semantic integration and found that effects of integration still appeared. Experiment 5 provided a stronger test, by crossing argumenthood with semantic integration. The experiment examined three different structures containing a local noun, as shown in (7).
(7) a. The report of the nasty auto accident
b. The report of the nasty auto accidents
c. The report that described the traffic accident
d. The report that described the traffic accidents
e. The report that Megan described the accident
f. The report that Megan described the accidents

The local noun (*accident(s)*) was part of an of-PP in cases like (7a and 7b), similar to the representational conditions in Experiment 1, and this served as a PP construction baseline. In cases like (7c and 7d), the local noun was in the object position of an RC (*that described the traffic accident(s)*), and in cases like (7e and 7f) the local noun was in the object position of an SC (*Megan described the accident(s)*). This set of conditions thus replicates Bock and Cutting’s (1992) experiments with singular-head items, except that (1) Bock and Cutting did not directly compare RCs and SCs (each was compared to PPs in separate experiments and showed a smaller head-local mismatch error effect), and (2) their PP conditions contained a variety of prepositions and were not controlled for semantic integration.

In order to determine whether Bock and Cutting’s clause packaging hypothesis (or hierarchical feature-passing) can account for results when semantic integration is controlled, the stimuli were selected, following plausibility and semantic integration norming, so that the of-PPs and RCs did not differ in mean degree of semantic integration. If clause packaging or hierarchical feature-passing has an effect in addition to semantic integration, the of-PPs and RCs should differ, with a larger mismatch error effect for the of-PPs.\(^6\)

The stronger test of argumenthood accounts versus semantic integration comes from a comparison of the RC and SC conditions. RCs are linked with their head nouns by way of coindexation — in (7c and 7d), for example, the report is what is describing the accident — and this should result in the head and local nouns being relatively closely integrated. SCs, on the other hand, are complete separate propositions, and thus nouns within them are semantically independent of the head noun and should not be tightly linked with it. If the effect of semantic integration for embedded clause constructions is like that in Experiments 1–4 for PP constructions, the RCs in this experiment should show a larger head-local mismatch error effect than the SCs. On the other hand, RCs adjoin to their head NPs, whereas SCs are typically treated as arguments of the head
noun. For argumenthood status to be responsible for the error effect interaction in Experiments 1 and 2 (of-PPs vs. with-PPs), local NPs within arguments of the head noun would have to create larger mismatch error effects than local NPs within adjuncts to the head’s NP. This predicts that the SCs should yield larger mismatch effects than the RCs in this experiment.

Method

Participants. Sixty Northeastern University undergraduates participated. One participant was dropped from the analysis due to a recording malfunction.

Materials and Design. Thirty-six stimulus sets like that shown in (7) were constructed. Each preamble consisted of a head NP followed by an of-PP (7a, 7b), an RC (7c, 7d), or an SC (7e, 7f). As in Experiments 1–4, all preambles ended with a singular or plural local noun. The head NP was always singular, and the six different versions of an item were created by varying the local NP number and the structure of the following modifier (of-PP vs. RC vs. SC). So that all versions in a stimulus set had the same number of syllables, extra words (most often adjectives) were inserted in the RC and the of-PP versions of the items.

In addition to the experimental items, 64 filler preambles were constructed, 26 of which had plural head NPs and 20 of which contained either a locative preposition (behind, beside, or near) or the preposition with. The 26 plural head items had structures like the experimental stimuli (a head NP followed by an of-PP, an RC, or an SC); the 20 locative and with-PP items had an NP PP structure. The experimental and filler stimuli were combined to form 6 counterbalanced 100-item lists. Each list included all 64 fillers and exactly one version of each of the 36 experimental preambles.

Semantic integration rating. The mean semantic integration ratings for the Experiment 5 stimuli are shown in Table 10. A 3 (structure) \times 2 (local noun number) ANOVA revealed no effect of local noun number and no interaction ($F$s < 1), but there was an effect of structure ($F(2, 70) = 37.90$, $MSe = .22$, $p < .001$). Participants rated the SC versions as less integrated than both the of-PP cases ($F(1, 35) = 51.87$, $MSe = .22$, $p < .001$) and the RC cases ($F(1, 35) = 62.17$, $MSe = .22$, $p < .001$), but there was no difference between the of-PPs and the RCs ($F < 1$).
Plausibility norming. Forty-nine Northeastern University undergraduates rated the preambles for plausibility. Fifty-one items with the same structures as those in (7) were used to create 6 counterbalanced lists with 25 or 26 items per page. Each list had a different random ordering of items. Participants were given the same rating scale and instructions as in the Experiment 1–4 plausibility surveys.

Thirty-six of the 51 rated items were chosen as the experimental stimuli for the experiment. The mean plausibility ratings of these items are shown in Table 10. An ANOVA on the ratings showed a main effect of structure ($F(2, 70) = 11.43, MS_e = .50, p < .001$): The RCs and the of-PPs were both rated as more plausible than the SCs (RC vs SC: $F(1, 35) = 21.36, MS_e = .24, p < .001$; of-PP vs. SC: $F(1, 35) = 9.82, MS_e = .32, p < .01$). The RCs and of-PPs did not differ from each other ($F(1, 35) = 1.30, MS_e = .18, p > .20$). There was no effect of local noun number and no interaction ($Fs < 3, ps > .10$).

Procedure and Scoring. Individual trials proceeded as in Experiments 1–4, except that preambles were presented for the larger of 1000 ms or 50 ms per character, rather than 40 ms per character. The same scoring procedure was used as in Experiments 1–4.

Results

The percentage of agreement error responses in each condition and the counts of each response type are shown in Table 11. Speakers produced significantly more errors in the plural local noun conditions than in the singular local noun conditions ($F_1(1, 58) = 20.66, MS_e = 369.25, p < .001$; $F_2(1, 35) = 26.49, MS_e = 174.49, p < .001$), and there was also a main effect of structure ($F_1(2, 116) = 12.46, MS_e = 171.41, p < .001$; $F_2(2, 70) = 12.83, MS_e = 124.23, p < .001$), with all 3 types differing from each other: Of-PP structures created the most errors (vs. RC: $F_1(1, 58) = 6.94, MS_e = 130.92, p < .05$; $F_2(1, 35) = 5.66, MS_e = 167.81, p < .05$; vs. SC: $F_1(1, 58) = 17.63, MS_e = 241.83, p < .001$; $F_2(1, 35) = 21.13, MS_e = 150.44, p < .001$), and RC structures generated more
errors than SCs ($F_1(1, 58) = 8.73, MS_e = 141.48, p < .01; F_2(1, 35) = 11.99, MS_e = 54.45, p < .01) 

In addition, structure and local noun number interacted ($F_1(2, 116) = 21.13, MS_e = 147.72, p < .001; F_2(2, 70) = 18.41, MS_e = 121.38, p < .001), with the largest head-local mismatch effect occurring for the of-PPs ($F_1(1, 58) = 27.68, MS_e = 372.66, p < .001; F_2(1, 35) = 27.73, MS_e = 276.10, p < .001), no mismatch effect occurring for the SCs ($F_1(1, 58) = 1.69, MS_e = 50.02, p > .15; F_2(1, 35) = 1.60, MS_e = 30.37, p > .20), and an intermediate effect for the RCs ($F_1(1, 58) = 14.34, MS_e = 242.02, p < .001; F_2(1, 35) = 12.50, MS_e = 110.79, p < .01). All the 2 (structure) x 2 (local NP number) ANOVA interactions comparing these effects were reliable, indicating that the sizes of the head-local mismatch effects differed for each structure (of-PP vs. RC: $F_1(1, 58) = 8.90, MS_e = 102.18, p < .01; F_2(1, 35) = 8.60, MS_e = 156.94, p < .01; of-PP vs. SC: $F_1(1, 58) = 29.27, MS_e = 209.59, p < .001; F_2(1, 35) = 32.91, MS_e = 135.60, p < .001; SC vs. RC: $F_1(1, 58) = 17.65, MS_e = 131.40, p < .001; F_2(1, 35) = 11.97, MS_e = 81.60, p < .01). These patterns did not change if the number of errors (rather than percentages) was analyzed, except that the of-PP versus RC main effect was not significant, and the of-PP versus RC interaction with local noun number was marginal by items.

More uninflected responses were produced following of-PPs than either RCs or SCs, yielding a main effect of structure ($F_1(2, 116) = 9.18, MS_e = .98, p < .001; F_2(2, 70) = 7.66, MS_e = 1.93, p < .01). However, there was no main effect of local noun number for uninflected responses, nor was there an interaction ($Fs < 1$).

For miscellaneous responses, there was an effect of structure ($F_1(2, 116) = 11.46, MS_e = .787, p < .001; F_2(2, 70) = 9.48, MS_e = 1.56, p < .001), with more miscellaneous responses occurring after SCs than after RCs or of-PPs. There was also an effect of local noun number ($F_1(1, 58) = 10.07, MS_e = .673, p < .01; F_2(1, 35) = 5.83, MS_e = 1.91, p < .05), with more miscellaneous responses following plural local nouns than singular local nouns. However, structure and local noun number did not interact ($Fs < 2, ps > .30$).
Discussion

The central result of Experiment 5 was that the head-local mismatch error effect was greatest for the of-PP constructions, next largest for the RCs, and smallest for the SCs, replicating and extending Bock and Cutting’s (1992) findings that local NPs within embedded clause structures create smaller error effects than those within PP structures. This pattern has implications for a variety of possible explanations of head-local mismatch error effects.

First, as with Experiments 1–4, the RC versus SC difference is predicted by a parallel activation-based system in which greater semantic integration results in greater interference between elements of phrases being planned and an increased head-local mismatch error effect. The difference is opposite that predicted by a memory-shift-based system.

Second, the difference between RCs and SCs reinforces the conclusions from Experiments 3 and 4, that argumenthood status of the structure containing the local NP does not have an influence on subject-verb agreement processes. If it did, the argument cases from this experiment (the SCs) and from Experiments 1 and 2 (the of-PPs) should have all yielded larger error effects than the corresponding adjunct comparison cases (RCs in this experiment, with-PPs in Experiments 1 and 2).

Third, the larger error effect for RCs than for SCs also cannot be accounted for by clause packaging, unless it is augmented with some means of distinguishing among clause types, such as sensitivity to coindexation or the presence of a relative pronoun (which does have to agree on at least some features with the coindexed head noun). These alternatives have never been proposed in this form, but nothing rules them out. Bock and Cutting (1992, p. 110) did note the coindexation and relative pronoun difference between RCs and SCs but suggested that the effect of the relative pronoun would be to reinforce the number of the singular head noun, which predicts (incorrectly) that RCs should yield a lower mismatch error effect than SCs. Thus if the existence of coindexation and the relative pronoun is to account for the RC versus SC difference, a more likely explanation is that they are responsible for the difference in semantic integration between the two cases, which in turn yields a difference in error effects.

Another possibility is that the singular number of the subject NP within the SC itself (e.g., Megan in (7e) and (7f)) created the RC versus SC difference, by reducing the probability of plural agreement in the SC conditions. This alternative account is problematic, however, for three reasons: First, while Bock and Cutting (1992) did find an influence of the number of the SC’s subject NP on
agreement error rates, with plural SC subjects increasing the rate of plural verb (error) responses, the error rate for their plural SC subject NP cases was still quite small (approximately 5%), while the error rate for the Experiment 5 RCs was approximately 10%. Second, Bock and Cutting found a similar increase due to a plural SC subject NP for both singular and plural local nouns, indicating that the effect of the embedded clause subject’s number is independent of the local noun number. Thus the Experiment 5 head-local mismatch error effect for the SC conditions, if they had had plural SC subject NPs, would be expected to remain very close to 0% (or negative), and clearly smaller than that for the RCs. Third, the singular SC subject NPs match the number of the RC subject NPs. Although the RCs did not have overt subject NPs, the implied subject was always the head noun of the main clause subject NP (e.g., report in (7c) and (7d)), which was always singular. Thus the presence of a singular subject NP inside the SC cannot explain the RC versus SC difference.

Whether hierarchical accounts can explain the RC versus SC difference depends, as in earlier experiments, on assumptions about what the target is for feature-passing (the subject NP vs. the head noun), which nodes count for distance (all vs. maximal projections only), and how RCs and SCs attach into the subject NP (see Figure 1, replacing the PP with the RC or SC; RCs can attach as in Figure 1B or 1C, and SCs can attach as in Figure 1A or 1C). Substituting SCs for the of-PPs and RCs for the with-PPs in Table 1 shows the possible predictions. Hierarchical feature-passing makes the correct prediction only if the target of feature-passing is the subject NP, all nodes are counted for distance, and SCs attach as in Figure 1A. Of course, this is just the set of assumptions which makes the wrong prediction for Experiments 1 and 2. Hierarchical accounts’ predictions also potentially depend on whether the path along which an errant feature must travel is different within the RC versus the SC structure. However, while the internal structure of RCs is a matter of debate, for the Experiment 5 stimuli most theories predict either no difference in path length (e.g., Pollard & Sag, 1994) or that RCs contain additional nodes (e.g., Chomsky, 1995). Neither of these cases creates an effect in the correct direction.7

The other main implication of the Experiment 5 result pattern is that some factor in addition to semantic integration is needed to account for the difference in error rates between the of-PP and RC conditions, which did not differ in rated semantic integration. Clause packaging and hierarchical feature-passing accounts will predict correctly that the of-PPs should yield a higher
mismatch error effect than the RCs. Because clause packaging does not predict a difference in all the other cases so far discussed, and semantic integration does not predict a difference for the of-PPs and RCs, a combination of semantic integration and clause packaging can explain the full range of Experiment 1–5 error patterns.

Combining semantic integration with hierarchical feature-passing may also be able to explain the full range of results, depending on assumptions noted in the previous experiments. The relevant cases are those in which the of-PPs and with-PPs are predicted to be equal in Table 1. One of the possibilities is that (a) distance is measured from the local noun to the subject NP (not the head noun), and (b) only maximal projections are counted. In this case hierarchical distance will not differentiate among any of the Experiment 1–4 conditions, nor will it differentiate the RC and SC conditions in Experiment 5, and semantic integration will make the appropriate prediction in all of those cases. Hierarchical distance will then account for the of-PP versus RC difference: Regardless of whether the two structures attach as in Figure 1A, 1B, or 1C, distance from the attachment point (the PP in Figure 1) will be the same, but the RC will contain at least one additional internal maximal projection — the VP — which must be traversed. Two other more specific combinations of assumptions would also create the appropriate predictions; these are the other two cases in Table 1 in which the of-PPs and with-PPs are predicted to be the same. A fourth possibility permitting hierarchical distance to combine correctly with semantic integration is that all of the phrases and clauses containing the local nouns in Experiments 1–5 attach into the subject NP in the same way. If this were true, the only hierarchical distance manipulations in any of the Experiment 1–5 stimuli would be for the of-PPs versus the RCs and SCs, so semantic integration would create the correct prediction for all of the other cases, and hierarchical distance would yield a shorter distance in the of-PPs than in the RCs. Whether any of these combinations of assumptions is correct is not clear.

Semantic Integration Meta-Analyses

Experiments 1–5 show that differences in degree of semantic integration predict differences in head-local mismatch error effects, and these results implicate a planning mechanism which tends to plan the elements of more integrated phrases more simultaneously. In addition to the effects of
integration within each experiment, however, there are differences across the experiments in degree of integration of the stimuli, and we therefore wished to determine how well semantic integration could predict error rates across the full set of conditions from all the experiments. To this end, we correlated plural local noun condition integration ratings with head-local mismatch error effects (plural local noun error rate — singular local noun error rate). We also examined the relationship between semantic integration and plausibility, because we wanted to insure that the former was not simply reducible to the latter. It would not be surprising to find a correlation between the two — when a phrase describes a less plausible situation, computing relationships between elements of the phrase is likely to be more difficult, leading to lower integration ratings — but if integration influences planning processes as suggested, then a relationship between integration and error rates should hold even after plausibility has been factored in.

In addition, we conducted a separate set of correlations to determine how well integration differences could account for error effects from a range of studies in the subject-verb agreement literature. Along with local noun number, these studies manipulated distributivity (Bock & Miller, 1991; Eberhard, 1999), local phrase length and construction type (Bock & Cutting, 1992; Bock & Miller, 1991), animacy and concreteness (Bock & Miller, 1991), local noun morphology (Bock & Eberhard, 1993), and head noun semantic properties (Bock et al., 1999); but none of them were designed to manipulate semantic integration. Thus these correlations should provide a more general test of the degree to which error effects are influenced by degree of integration. We also used these results to examine whether integration was confounded with any of these other factors. We were particularly interested in distributivity, because like semantic integration, its effects are created from the message-level representation used in planning. Furthermore, while it might turn out that integration can replace distributivity in cases where the latter has been claimed to influence number-agreement error rates (Eberhard, 1999; cf. Bock & Miller, 1991), distributivity cannot replace integration in our experiments: In Experiment 1, for example, the versions containing with-PPs seem more likely than those containing of-PPs to be conceptualized as multiple-token (the drawing with the flowers involves multiple distinct objects, whereas the drawing of the flowers involves only a single object), but the with-PPs yielded smaller error effects than the of-PPs. The Experiment 4 stimuli are similar: The accompaniment versions seem more likely to involve multiple tokens than the attribute versions, but the former yielded smaller error effects. And within each of
Experiments 2, 3, and 5, the conditions do not seem to differ in this property.

Finally, we conducted a third set of analyses\(^8\), to examine whether semantic integration might be confounded with another factor, plural markedness, which has been examined in some detail in the subject-verb agreement literature. The repeated finding that head-local mismatch effects do not appear with preambles headed by plural nouns (e.g., Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991; Bock et al., 1999; Eberhard, 1997; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco et al., 1995; cf. Franck et al., 2002; Vigliocco & Nicol, 1994; also in comprehension: Nicol et al., 1997; Pearlman, Garnsey, & Bock, 1999; cf. Pearlman, 2000) has been taken as evidence that the plural form is marked with an explicit feature which triggers plural agreement, while the singular form is the default and does not typically have an explicit feature (e.g., Eberhard, 1997). However, given the results of Experiments 1–5, an alternative account of the lack of mismatch effects for plural head noun constructions is that such cases are typically not as semantically integrated as singular head noun constructions. None of our experiments compared singular-head and corresponding plural-head stimuli, but this account seems possible a priori, at least in the sense that plural head nouns generally require multiple conceptual tokens, which might in turn necessitate a lower degree of integration relative to a single conceptual token (e.g., *three keys* might have an intrinsically less integrated conceptual representation than *one key*). Of course, this may not influence the degree to which the head noun and local noun are integrated, but to examine this possibility, we compared semantic integration ratings for singular-head and corresponding plural-head stimuli from prior studies. If the plural-head stimuli have consistently lower integration ratings, then semantic integration differences might provide an alternative account of the results taken as evidence for plural markedness.

**Method**

The first set of analyses was conducted using the error rate results from Experiments 1–5 and the semantic integration and plausibility rating data collected for those experiments. The second set of analyses was conducted using the error and correct agreement count data from the studies in the literature. Error rates in each condition were computed as in Experiments 1–5. Semantic integration ratings for the stimuli from prior studies were obtained in two new surveys similar to that described in Experiment 1. The first of these (Survey 1) included only singular-head versions of
the stimuli, and these ratings were compared to mismatch error rates in the second set of analyses, and to ratings from Survey 2 (which contained primarily plural-head versions) in the third set of analyses.

**Participants.** Eighty-one participants rated the items from prior studies in Survey 1. Another 80 participants completed Survey 2. All were Northeastern University undergraduates.

**Materials and Design.** Table 12 shows the studies and specific experiments and conditions which provided stimuli for the second and third meta-analyses. Given the number of studies in the literature, integration ratings could not be obtained for all of their stimuli, so we focused on studies (1) using English materials, (2) for which the stimuli and appropriate error rate data had been published or were otherwise conveniently available, (3) which compared match and mismatch conditions with singular-head stimuli, and (4) which contained additional manipulations of potential interest (distributivity, etc., as noted above).

In Survey 1, only singular-head versions of stimuli were included, and ratings were always obtained for both the singular and plural local noun versions. There were 160 items in total, each with multiple versions: the 16 PP items from Bock and Miller (1991, Experiment 1), manipulating preamble length and distributivity (e.g., *The label on the (tamper-proof medicine) bottles, The key to the (ornate Victorian) cabinets*); the 32 items from Bock and Miller’s Experiment 2, manipulating head noun animacy (e.g., *The author of the speeches, The speech of the authors*); the 32 Bock and Cutting (1992, Experiment 1) items, comparing PPs to RCs (e.g., *The editor of the history books, The editor who rejected the books*); the 32 Bock and Cutting Experiment 3 items, comparing PPs to SCs and manipulating preamble length (e.g., *The report of the destructive (forest) fires, The report that they controlled the (forest) fires*); the 36 verb-agreement items from Bock et al. (1999), comparing singular and collective heads (e.g., *The actor/cast in the soap operas*); and 12 items from Eberhard (1999, Experiment 2), manipulating distributivity (e.g., *The name on the credit cards,*
The check from the stockbrokers) (the other 4 items in this experiment were also included separately, as part of the Bock & Miller Experiment 1 stimuli). The 32 Bock and Miller (1991, Experiment 2) items also included a manipulation of relative concreteness of the head and local nouns (head more concrete than local vs. head less concrete than local), and we computed semantic integration means for these two conditions separately. However, the analyses relating semantic integration to error rates included only the manipulation of animacy, because Bock and Miller's original Experiment 2 design contained a counterbalancing error, and the data from their replication, which corrected the error, were not available.

Eight lists were constructed for Survey 1 such that no list contained more than one version of any item. Each list contained 96 stimuli, with 16 items per page, and each version of each item was rated by approximately 10 participants. Items were rated on a scale from 1 (not linked) to 7 (tightly linked), as in the Experiment 1 integration survey, and all other aspects, including the instructions and randomization of pages, were identical to those in Experiment 1.

Survey 2 contained primarily plural-head versions of the same items and conditions as Survey 1, as shown in Table 12, except that (1) the 32 Bock and Cutting (1992, Experiment 3) items were included as singular-head versions for direct comparison with Survey 1 (Bock & Cutting did not use plural-head versions of these stimuli); (2) the 12 items from Eberhard (1999, Experiment 2) were not included; and (3) the 8 items from Bock and Eberhard (1993, Experiment 3) were included, manipulating head noun number, local noun number, and local noun plural morphology (e.g., The trap(s) for the mouse/mice/rat/rats). These changes yielded 156 items, each with multiple versions, which were counterbalanced in 8 lists, as in Survey 1. Each list contained either 88 or 90 items, spread over 6 pages. All other aspects of Survey 2 were identical to Survey 1.

Results and Discussion

The dependent variable for the correlations and regressions in the first two meta-analyses was the head-local mismatch error effect (plural local noun – singular local noun), but the patterns did not change if the error rate from the plural local noun condition alone was used instead.

Relationship between semantic integration, plausibility, and errors across Experiments 1–5. Figure 2 shows the central result of the first meta-analysis, relating head-local mis-
match error effect to plural local noun condition semantic integration rating for the 12 pairs of conditions from Experiments 1–5. This correlation was reliable \( r = .77, p < .01 \), reinforcing the conclusions from the factorial analyses in Experiments 1–5: Error effects increase with degree of integration.

=================================
Insert Figure 2 About Here
=================================

We also examined how plausibility might interact with this result. Semantic integration and plausibility were reliably correlated across the 12 plural local noun conditions \( r = .68, p < .05 \), suggesting that plausibility might replace integration. However, the correlation between mismatch error effect and plausibility was not reliable \( r = .47, p > .10 \), and in a multiple regression analysis with plausibility and semantic integration as independent variables \( R^2 = .60, p < .05 \), only the coefficient associated with semantic integration was reliable \( \beta = .84, p < .05 \); plausibility \( \beta = -.10, p > .70 \). As a final check, we also conducted two stepwise hierarchical regression analyses with these variables \( F\text{-to-enter} = 4 \). In the first, plausibility was forced in as the first factor, but semantic integration still entered at the next step to yield a substantial incremental change in \( R^2 \) (.37, \( p < .05 \)). In the second hierarchical analysis, semantic integration was forced in first, and in this case, plausibility contributed nothing further \( F\text{-to-enter} < 1 \). Thus while plausibility and degree of semantic integration are related, plausibility differences cannot explain our results.

**Effects of semantic integration in stimuli from prior studies.** The second meta-analysis examined whether semantic integration could predict mismatch error effects in singular-head conditions of studies not intended to manipulate it (see Table 12). Table 13 shows the mean ratings by condition from Survey 1 for the stimuli from prior studies, and Figure 3 illustrates the correlation between the singular-head mismatch ratings and corresponding mismatch error effects. (As noted above, the relative concreteness factor manipulated in Bock and Miller’s (1991) Experiment 2 was excluded from the correlational analysis shown in Figure 3.) Two additional data points, the singular-head regular and irregular local noun morphology condition pairs from Bock and Eberhard (1993, Experiment 3), were included in this analysis as well. The ratings for these conditions were
collected in Survey 2, but they were otherwise comparable to the data from the other prior studies. As with Experiments 1–5, semantic integration was reliably correlated with head-local mismatch effects across the 18 singular/plural local noun condition pairs examined ($r = .75$, $p < .001$). Despite the fact that the stimuli from prior studies had not been designed to manipulate semantic integration (and had been designed to manipulate a variety of other factors), the magnitude of the correlation was nearly identical to that for Experiments 1–5, as were the slopes and intercepts of the corresponding regression lines (slopes: 9.8 vs. 8.8; intercepts: $-25.1$ vs. $-20.1$ for Experiments 1–5 and the stimuli from prior studies, respectively).

Insert Table 13 About Here

Insert Figure 3 About Here

We also examined several of the specific contrasts among conditions from prior studies (see Table 12), to determine whether semantic integration differences might provide an alternative explanation. For these analyses, we conducted separate ANOVAs for each experiment from the prior studies, using factors in Table 12; depending on the study, the factors were one or more of: distributivity, length, head noun semantic number (singular vs. collective), and construction (PP vs. RC vs. SC). We did not examine head noun animacy, relative head/local concreteness, or local noun plural morphology in detail, because these factors did not have effects on error rates in the literature (though cf. Barker et al., 2001, for some discussion of animacy, and Haskell & MacDonald, 2001, for morphophonology). Each ANOVA had items as the random factor and included only the singular-head mismatch condition integration ratings (see Table 13) as the dependent variable. The patterns of results from these ANOVAs were then compared to the head-local mismatch error effects reported in the studies (see Figure 3).

We first examined distributivity. Eberhard (1999) found that distributive (multiple-token)
preambles yielded larger head-local mismatch error effects in English. This was the case for her Experiment 2 stimuli, and their integration ratings reflected the difference in error rates: Eberhard’s distributive stimuli were marginally more integrated than her nondistributive stimuli \( (F(1, 14) = 4.03, MS_e = .81) \). Bock and Miller (1991, Experiment 1), on the other hand, did not find an effect of distributivity on error rates. Eberhard showed (using ratings) that the Bock and Miller stimuli were less imageable than her own and argued that sufficient imageability was a prerequisite for a conceptually plural message-level representation to influence grammatical agreement. The integration ratings for the Bock and Miller Experiment 1 stimuli provide an alternative explanation, in that the distributive and nondistributive items did not differ in degree of integration \( (F < 1) \). Thus both the Eberhard (1999) effect of distributivity and the Bock and Miller (1991) lack of an effect of distributivity might be explained by semantic integration instead. Sufficient imageability might still be a prerequisite for effects of integration to appear (in either ratings or error effects), but imageability cannot replace integration (e.g., Eberhard’s distributive stimuli were not more imageable than her nondistributive stimuli) nor can distributivity — as noted above, it cannot explain the results of our Experiments 1–5.

We next examined length effects in the prior studies. Bock and Cutting (1992, Experiment 3) found that longer PPs intervening between the head noun and the number-marked verb yielded larger head-local mismatch error effects, but this pattern of results did not appear in the integration ratings for these items: The ANOVA yielded no effect or interaction involving length \( (Fs < 1) \). Bock and Miller (1991, Experiment 1) also examined length effects, and the integration rating ANOVA for these stimuli also revealed no effect of length \( (F < 1) \) and no interaction (with distributivity, \( F(1, 14) = 2.94, MS_e = .39, p > .10 \)). Bock and Miller in fact found no reliable effect of length on error rates for their Experiment 1 stimuli, but the reported comparison included both PP and RC preambles (as in our Experiment 5), and length effects may not occur for preambles with intervening clauses (cf. Bock & Cutting, 1992, Experiment 3). In any case, however, there was no evidence that differences in semantic integration could account for effects of preamble length.

Semantic integration also could not replace the head noun singular versus collective semantic property. Bock et al. (1999) showed that collective-head error rates were higher than singular-head error rates, but this difference was not reflected in integration ratings \( (F < 1) \).

Finally, we examined whether the effects of preamble construction (whether the local noun was
part of a PP, RC, or SC) in Bock and Cutting (1992, Experiments 1 and 3) followed the same pattern as in our Experiment 5, which showed that semantic integration could account for the difference in error rates between RC and SC constructions, but that an additional factor (e.g., clause packaging) is needed to explain why error rates were lower for RC constructions than for corresponding PP constructions. The integration rating ANOVAs for the Bock and Cutting stimuli revealed an identical pattern: For their Experiment 1 stimuli, the PP and RC integration ratings did not differ \((F < 1)\), despite the difference in error rates they found; but for their Experiment 3 stimuli, the PP versions were more integrated than the SC versions \((F(1, 31) = 4.34, MS_e = .66, p < .05)\). An additional comparison of their Experiment 1 RC items and their Experiment 3 SC items revealed that the RCs were more integrated \((F(1, 62) = 103.3, MS_e = .69, p < .001)\). Thus as for our Experiment 5, a combination of semantic integration and a factor like clause packaging or hierarchical feature-passing is needed.

**Semantic integration as a replacement for plural markedness.** The third meta-analysis compared semantic integration ratings for singular-head and corresponding plural-head stimuli from the prior studies, in order to determine whether semantic integration might provide an alternative explanation to plural markedness for the lack of local noun mismatch effects with plural-head stimuli. The relevant singular-head and plural-head mismatch condition means are shown in Table 13. For an integration-based explanation to replace plural markedness, singular-head mismatch stimuli would have to have higher integration ratings than corresponding plural-head mismatch stimuli. As is apparent from Table 13, this was not the case in general; and a 5 (study) \(\times\) 2 (head noun number) ANOVA with items as the random factor supported this conclusion: Stimuli with singular head nouns were in fact reliably *less* integrated than corresponding versions with plural head nouns \((F(1, 119) = 4.37, MS_e = .18, p < .05)\), and an interaction with study was also present \((F(4, 119) = 5.80, MS_e = .18, p < .001)\). The combination of the main effect of head noun number and the interaction makes clear that semantic integration cannot provide an alternative explanation for effects which have been attributed to plural markedness.
General Discussion

The central result of Experiments 1–5 and the first two meta-analyses is that semantic integration, the degree to which two parts of a message are linked within a conceptual representation, has a consistent and substantial influence on subject-verb agreement processes: Head-local mismatch effects in the production of subject-verb agreement are larger for cases that are more semantically integrated. This was true for the of-PP versus with-PP cases in Experiments 1 and 2, for the for-PP versus with-PP cases in Experiment 3, for the attribute versus accompaniment PPs in Experiment 4, and for the of-PP and RC cases versus the SCs in Experiment 5. Semantic integration was also strongly correlated with error effect magnitude across all these conditions, and across a variety of conditions from the subject-verb agreement error literature. This pattern of results could not be explained by variations in subject NP plausibility, clause packaging (Bock & Cutting, 1992), hierarchical feature-passing (Bock et al., 2002; Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994, 1998), the argument/adjunct distinction, distributivity or other variation in subject NP conceptual number (e.g., Bock et al., 1999, 2002; Bock & Miller, 1991; Eberhard, 1999; Vigliocco et al., 1995; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Hartsuiker, et al., 1996), or properties specific to particular nouns or prepositions.

The effect of semantic integration on agreement has important implications for how the language production system coordinates syntactic planning processes: It provides evidence against a syntactic planning system which is serial, relying on memory-shifting operations to change the focus of processing. Such a system should yield lower error rates in cases which are more integrated, because memory-shifting processes should be less resource-intensive in those cases. The results are instead as predicted by a parallel activation-based planning system in which some of the primary costs of planning arise in maintaining multiple active phrases simultaneously while minimizing interference between them. This class of planning mechanism should yield higher error rates in cases which are more integrated, because more integrated phrases will tend to be active simultaneously during processing and will thus be more likely to interfere with each other. In the case of subject-verb agreement, one manifestation of this interference will occur when properties (e.g., number) of the local noun conflict with those of the head noun, and the properties of the local noun inappropriately influence agreement. Whether this influence arises in computing the number of the subject NP or, instead, during the computation of the verb’s number on the basis of the subject NP’s value remains
an open question.

Although the effect of semantic integration on agreement error rates is clear, we have not provided any direct evidence that it arises through increased simultaneity in particular (though see Nicol, 1995, for a related proposal based on timing of activation). One alternative might be that greater semantic integration increases the overall activation level of constituent phrases, rather than altering the timecourse of their activation, similar to the effect Eberhard (1999) imputed to concreteness/imageability. The current results do not rule out such a mechanism, but it seems less likely than the simultaneity explanation, for two main reasons: First, this mechanism should have the same effect on the head noun’s representation as on the local noun’s representation — semantic integration is a symmetric relation for a given pair of elements in a particular construction — so the head noun should be better able to withstand increased interference from the local noun. Indeed, depending on exactly how subject-verb agreement is implemented, increased activation of the head noun might make subject-verb agreement errors even less likely, rather than more likely. In principle, though, this concern could be overcome with additional assumptions about non-linearity of the activation function (e.g., the head noun might already be close to its maximum at the relevant point in planning and thus less influenced by an additional boost) or about an interaction between the effect of integration and plural markedness (e.g., Eberhard, 1997; the plural feature on the local noun might be more affected than the (possibly absent) singular feature on the head).

The second reason for preferring the simultaneity-based explanation for the effect of semantic integration is that it appears to be more compatible with other production results. For example, sequencing errors such as anticipations (e.g., cup of coffee uttered instead as cuff of coffee, Fromkin, 1971) are typically modeled within parallel activation-based systems as cases in which two elements that might be placed in a particular position are active simultaneously, with the planning system selecting the wrong one (see Dell, Burger, & Svec, 1997, and references therein). Increasing the activation of the incorrect element without increasing its temporal overlap with the correct element will not have the same effect. Given the current results, increased simultaneity can thus be responsible both for certain sequencing errors and for some cases of agreement errors, depending on the situation and possibly also on the degree of simultaneity (e.g., agreement errors might begin to appear at a lower overlap threshold). Simultaneity is also directly connected to effects of timing of activation on word order choice in syntactic planning: Bock (1986, 1987b) found that words which
are semantically or phonologically primed (and which are therefore activated more quickly) tend to be produced earlier in an utterance if the message to be conveyed can be expressed with multiple alternative orderings, and tend to be produced with fewer hesitations and dysfluencies. Similarly, V. Ferreira (1996) showed that utterances are more fluent and are produced more quickly when multiple ordering options are available, a result predicted by models which can select among alternative plans based on whichever words become available (activated) first. The interesting difference between sequencing errors and the current agreement error results, on one hand, and Bock’s and Ferreira’s results, on the other, is that in the latter, flexibility in the sequencing plan is available (courtesy of the grammar), and the planning system makes use of it to generate an utterance with increased fluency. When, as in the former cases, flexibility in sequencing is not available, the same underlying mechanism — variation in timing of activation — instead yields dysfluencies and errors.

Another implication of the experiments and meta-analyses is that, while semantic integration is an important determinant of the difficulty involved in computing subject-verb agreement successfully, it is not sufficient to explain the full range of effects on agreement error rates. Some other factor is needed to account for the Experiment 5 difference in error effects between the of-PPs and RCs, and semantic integration did not vary for several of the variables manipulated by studies included in the second and third meta-analyses (e.g., phrase length, head noun conceptual number, plural markedness). In the case of the Experiment 5 of-PP versus RC result (which also appeared in the second meta-analysis for Bock & Cutting’s (1992) Experiment 1 results), either clause packaging or hierarchical feature-passing seem to be the likely candidates, so while neither can account for the current results on its own, one or the other still seems to be needed. None of the current results decides between them, but the combination of hierarchical feature-passing with semantic integration would have an interesting additional implication for the nature of the syntactic planning system and agreement computation in particular: It would suggest that agreement is computed by hierarchical feature-passing as syntactic phrases are being built, such that two phrases which are tightly integrated will be active simultaneously, making it easier for the agreement features from one phrase to be (incorrectly) passed into the other. This is in contrast to the computation of agreement by hierarchical feature-passing performed on a completed syntactic structure. In this case, no effects of integration on agreement processing would be predicted, even if greater integration lead to increased simultaneity of planning during structure-building. The
combination of hierarchical feature-passing and integration would also be incompatible with an interference account of integration effects in which simply having the head and local nouns active more simultaneously results in direct interference between their number properties, as opposed to an interaction between their number properties by way of feature-passing.

The meta-analyses also suggested that semantic integration may be confounded with other factors considered in the subject-verb agreement literature, and thus it will need to be controlled in future studies intended to investigate other variables. The results of the second meta-analysis provided some relevant preliminary data on one such case, suggesting that effects (and the lack thereof) attributed to distributivity manipulations in the literature on English (e.g., Bock & Miller, 1991; Eberhard, 1999) might in fact instead be effects of semantic integration. Distributivity is of particular interest, because it may be the only factor influencing syntactic agreement processes so far investigated which is dependent on properties of the conceptual message and is also clearly non-lexical (see also Bock et al., 2001, for some discussion). Nearly all other effects are arguably directly attributable either to lexical properties linked directly to the head or local noun’s number-marking (conceptual number and plural morphophonology effects, Bock & Eberhard, 1993; Bock et al., 1999, 2001, 2002; Haskell & MacDonald, 2001; Vigliocco et al., 1995), to lexical properties involved in determining subjecthood (animacy and semantic overlap of the head and local nouns, Barker et al., 2001; Bock & Miller, 1991), to syntactic processes involved in computing agreement (phrase structure, length, and hierarchical distance effects, Bock & Cutting, 1992; Bock & Miller, 1991; Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994, 1998), or to an interaction between lexical properties and syntactic processes (plural markedness, e.g., Bock & Miller, 1991; Eberhard, 1997; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco et al., 1995).

Only two other relevant effects might be attributed to a direct influence of the conceptual (message-level) representation on agreement. The first of these is Humphreys and Bock’s (in press) finding that preambles headed by collective nouns yield larger mismatch error effects when the PP encourages the notionally plural interpretation of the collective compared to when it encourages the notionally singular interpretation (e.g., The gang on the motorcycles vs. The gang with the motorcycles). Humphreys and Bock report that their notionally plural stimuli were rated as less integrated than their notionally singular stimuli, indicating that semantic integration cannot account for their effect and suggesting that conceptual number can influence agreement. However, one alternative is
that the message-level construal of the collective phrase as notionally plural or singular has its effect on the lexical representation of the collective noun, which then influences agreement processes as in Bock et al.’s (1999) examination of preambles with collective versus individual heads: Because the collective heads in Humphreys and Bock’s stimuli can be construed as either singular or plural, their lexical specification for number should permit either a singular or plural value. During lexical selection of such a noun, the properties of the message might thus be able to set the noun’s number specification value (i.e., if the noun’s lexical entry does not strongly specify either singular or plural, the message’s conceptual number may shift the noun’s lexical specification). Then, during the agreement process (marking, in Bock et al.’s, 2002, model), the number specification from the head would determine the verb’s number-marking as usual, including the potential influence of the local noun. In this case, the effect of the message’s conceptual number on tracking of syntactic number would only be indirect, with no influence of the conceptual representation on syntactic processes.

The other effect potentially attributable to a direct influence of the message level on agreement processes is Thornton and MacDonald’s (2003) finding that under at least some conditions, mismatch error effects are larger when the local noun is a plausible subject for the agreeing verb phrase than when it is not (e.g., *The album by the composers... was praised* vs. *... was played*; cf. Barker et al., 2001). This result is not likely attributable to lexical influences, but in fact it may be explainable in terms of semantic integration: Participants were given the verb to use in passive voice with each preamble, and cases in which both the head and local nouns are plausible as subjects of the passive verb tend also to be cases in which both nouns are more tightly linked to the predicate (e.g., praising an album also suggests that its composers are being praised, while playing the album does not involve playing the composers). If greater integration between (elements of) the subject NP and the predicate affects planning like greater integration within the subject NP, planning of the VP (including number assignment) will be more likely to occur simultaneously with activation of number properties of the local noun, resulting in a greater chance of interference. Thus as in the Humphreys and Bock (in press) case, these results may not provide evidence for a direct influence of the message level on agreement independent of semantic integration. These alternative explanations obviously require further specification and support, but if distributivity effects turn out to be attributable instead to semantic integration differences, this would appear to undermine the only clear evidence in support of any direct influence of message-level representations on syntactic
agreement processes.

One final issue is the nature of semantic integration. We defined it as the degree to which two elements in a conceptual representation or mental model are linked together, where the strength of the link depends on the relationship specified by the content of the particular conceptual representation. Thus particular syntactic properties are likely to influence degree of integration, as, for example, the presence of a syntactic trace coindexed with the head noun in the subject position of the RCs in Experiment 5, or effects of hierarchical or linear distance (e.g., Franck et al., 2002). However, these effects should be limited to cases in which the syntactic properties have corresponding effects in the conceptual representation. Similarly, while semantic integration was correlated with plausibility for our stimuli, this need not be the case in general (cf. the bracelet made of silver and the bracelet made of graphite; these should be similar in integration but not plausibility), nor is it necessarily correlated with degree of lexical-semantic overlap or simple lexical association. As noted above, and perhaps most important for subject-verb agreement results, semantic integration is also not related to conceptual number. Thus the Experiment 1–5 results are not attributable to an effect of the content of the message level on the processing of corresponding content in the syntax (cf. distributivity). Instead, they depend directly on the underlying mechanism (activation, in a parallel activation-based system) used to represent the message and to link the elements of the message to their syntactic counterparts during the planning process.

This has (at least) two important consequences: First, semantic integration should apply more broadly than just to pairs of nouns within subject NPs, and it may have effects on processing that have nothing to do with agreement computations. As suggested above, integration relationships can be expected to hold between nouns and verbs just as between pairs of nouns, and, more generally, some degree of integration should in principle exist between any two elements of a message. In language production, at least, these elements might have to be concurrently present in the message, such that effects of integration are delimited by utterance boundaries or the extent of message-level planning units. Depending on the task and construction, the influence of integration might in turn appear as effects on error rates, production latencies, phrase ordering, or some other measure.

The second consequence of the nature of integration effects is that they appear to be closely related to effects in comprehension of local discourse cohesion (e.g., Halliday & Hasan, 1976) and the operations underlying identification and maintenance of anaphor-antecedent relationships (e.g.,
Dell, McKoon, & Ratcliff, 1983). These are cases in which potentially novel links (not necessarily dependent on lexically-based or plausibility-determined relationships) are formed within the discourse representation. In the case of anaphor-antecedent relations, as in semantic integration cases, the existence of a tight link then results in an increase in activation of one element when the other is active, as a function of the basic mechanisms involved in computing the discourse representation. Indeed, in comprehension, anaphor-antecedent relations might just be special cases of semantic integration, involving a very tight link (presumably the tightest possible). Of course, the effect of semantic integration in comprehension might not be the same as in production, given the different relationships between syntactic processes and discourse or message-level processes in the two cases. Nevertheless, we might expect semantic integration effects on the comprehension of subject-verb agreement, for example, analogous to those in production; and other effects of semantic integration might be measured fairly directly by probing discourse representations during ongoing processing. These kinds of investigations could be very useful in understanding the relationship between comprehension and production processes and the detailed nature of semantic integration and its effects on both syntactic planning and sentence and discourse comprehension.
References


Author Note

This research was supported by NSF grant BCS-9729007 and NIH grant 1 R01 DC05237-01 to NJP. We greatly appreciate the comments and advice of Bill Badecker, Kay Bock, Gary Dell, Kathy Eberhard, Vic Ferreira, Robert Frank, Karin Humphreys, Joanne Miller, Martin Pickering, Fei Xu, and an anonymous reviewer on earlier presentations of this work. We also thank Seth Amsden, Vened Argaman, Mathew Arruda, Joanna Currier, Angela Ficken, Grazyna Furman, Yuka Kominato, Joanne Malinka, Aurora Mendelsohn, Kimberly Mitchell, Joey Pasquino, Shelagh Scollin, Elizabeth Tatro, Naa-Solo Tettey, and Jennifer Turco for aid in conducting the experiments and coding and analyzing data. Portions of this work were presented at the 2001 CUNY Sentence Processing Conference (Philadelphia, PA), the 2001 Psychonomic Society Conference (Orlando, FL), and the 2002 CUNY Sentence Processing Conference (New York, NY).
Footnotes

1The issue of which mechanism implements syntactic planning is separate from the contrast between competitive and incremental models as discussed by V. Ferreira (1996; see also Stallings et al., 1998). The latter issue deals with how the planning system determines an ordering for constituents (phrases) when flexibility is available, and in particular whether possibilities compete versus whether the system constructs whichever alternative becomes available first; the former issue deals with how the syntactic planning system implements a particular ordering. These two parts of grammatical encoding will obviously be linked together, but none of the possible combinations is incompatible. We return to the relationship between these two issues in the general discussion.

2In prior work, the subject NP maximal projection has been the assumed target, because the subject NP’s marking can then be straightforwardly linked to the VP’s marking, in order to implement agreement on the verb. Thanks to a reviewer for pointing out this alternative.

3The complete list of stimuli for this experiment and Experiments 2–5, along with semantic integration and plausibility norms for each version of each item, is available for download at http://www.psych.neu.edu/pearlmutter/lab/papers.html.

4We also conducted paired comparisons of the percentage of error responses for the plural local noun conditions in each of Experiments 1–5, because one reviewer was concerned that the singular local noun conditions might not always be appropriate controls. These paired comparisons displayed statistical patterns identical to the corresponding ANOVA interactions (i.e., when the interaction was significant, the paired comparison using only the plural noun conditions was significant; when the interaction was marginal, the paired comparison was marginal), except for the Experiment 2 analysis by participants, in which the interaction was reliable but the paired comparison was marginal.

5This explanation is independent of one in terms of hierarchical feature-passing based on the potential structural consequences of the argument/adjunct distinction.

6In the case of hierarchical feature-passing, the prediction depends on the assumption that differences in distance associated with different attachment points (the positions of the PP in
Figure 1) for the of-PP versus the RC are smaller than the extra distance within the RC through which an errant feature must travel. However, the distance difference due to different attachments is at most just one syntactic node (e.g., if the of-PPs attach as in Figure 1A, the RCs attach as in Figure 1B or 1C, and the total number of nodes from the local NP to the subject NP maximal projection is the relevant distance measure). Thus as long as RCs contain at least two internal nodes through which an errant feature must pass (e.g., VP, IP; or VP, S), hierarchical feature-passing accounts will predict a higher mismatch error rate for the of-PPs than the RCs.

Another possibility is that hierarchical feature-passing might be sensitive to traces (e.g., Chomsky, 1981), as would be found in the subject position of the RCs, so that the relevant syntactic distance would be from the local noun to the subject NP of the RC. In this case, the syntactic distance would be shorter in the RC conditions than in the SC conditions, providing an alternative explanation for the higher error rate in the RC case. However, this account ignores the additional syntactic distance from the trace to the head NP; if the incorrect plural number feature is not passed along this additional distance, it will not be able to influence marking of the matrix clause verb, and will only have a potential effect on the RC verb. If the additional trace-to-head-NP distance is counted, the path is as long or longer in the RC conditions than in the SC conditions.

We are indebted to Kay Bock for suggesting these.

Because the singular-head and plural-head versions were mostly presented in separate surveys (Survey 1 and Survey 2, respectively), one concern is that the main effect of head noun number was the result of some confounded difference between the surveys. To control for this possibility, the singular-head versions of the Bock and Cutting (1992, Experiment 3) stimuli were included in both Survey 1 and Survey 2. A comparison between the ratings of these items in the two surveys indicated that they did differ reliably ($F(1,31) = 7.99, MS_e = .039, p < .01$), but the Survey 1 ratings were higher than the Survey 2 ratings (by .14). Correcting the Survey 2 ratings for this difference would only have increased the magnitude of the existing main effect of head noun number.
### Table 1: Alternative Hierarchical Feature-Passing Predictions

<table>
<thead>
<tr>
<th>Target Node</th>
<th>Of-PP Structure</th>
<th>With-PP Structure</th>
<th>Counted Nodes</th>
<th>Of-PP Path</th>
<th>With-PP Path</th>
<th>Error Rate Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject NP</td>
<td>A</td>
<td>B</td>
<td>Max</td>
<td>PP, NP</td>
<td>PP, NP</td>
<td>Of = With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>All</td>
<td>PP, N', NP</td>
<td>PP, NP</td>
<td>Of &lt; With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>Max</td>
<td>PP, NP</td>
<td>PP, NP</td>
<td>Of = With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>All</td>
<td>PP, N', NP</td>
<td>PP, NP</td>
<td>Of &lt; With</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>B</td>
<td>Max</td>
<td>PP, NP</td>
<td>PP, NP</td>
<td>Of = With</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>B</td>
<td>All</td>
<td>PP, NP</td>
<td>PP, NP</td>
<td>Of = With</td>
</tr>
<tr>
<td>Head Noun</td>
<td>A</td>
<td>B</td>
<td>Max</td>
<td>PP</td>
<td>PP, NP, NP</td>
<td>Of &gt; With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>All</td>
<td>PP, N', N</td>
<td>PP, NP, NP, N</td>
<td>Of &gt; With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>Max</td>
<td>PP</td>
<td>PP, NP</td>
<td>Of &gt; With</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>All</td>
<td>PP, N', N</td>
<td>PP, NP, N</td>
<td>Of = With</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>B</td>
<td>Max</td>
<td>PP, NP</td>
<td>PP, NP, NP</td>
<td>Of &gt; With</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>B</td>
<td>All</td>
<td>PP, NP, N</td>
<td>PP, NP, NP, N</td>
<td>Of &gt; With</td>
</tr>
</tbody>
</table>

*Note.* Letters for PP structures refer to Figure 1. Max = Maximal projections only. Paths are the sequences of nodes (see Figure 1) through which a feature must pass in order to create an error. This includes the PP node and the target of feature-passing (the subject NP or head noun).
Table 2: Experiment 1 Semantic Integration and Plausibility Ratings

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun</th>
<th>Semantic Integration</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of</td>
<td>Singular</td>
<td>4.13 (.93)</td>
<td>4.90 (1.10)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>4.16 (.92)</td>
<td>4.61 (1.12)</td>
</tr>
<tr>
<td>With</td>
<td>Singular</td>
<td>3.49 (1.03)</td>
<td>3.85 (.93)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>3.36 (1.09)</td>
<td>4.08 (.77)</td>
</tr>
</tbody>
</table>

Note. Rating scales were 1 (not linked/plausible) to 7 (tightly linked/plausible). Standard deviations are shown in parentheses.
Table 3: Experiment 1 Error Rates (%) and Response Counts by Condition

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun Number</th>
<th>Error Rate</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of Singular</td>
<td>1 (0.5, 1.4)</td>
<td>1 (0)</td>
<td>153 (21)</td>
<td>33 (3)</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Plural</td>
<td>21 (4.1, 4.4)</td>
<td>32 (7)</td>
<td>118 (14)</td>
<td>29 (2)</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>With Singular</td>
<td>0 (0.0, 0.0)</td>
<td>0 (1)</td>
<td>123 (22)</td>
<td>31 (1)</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Plural</td>
<td>7 (2.7, 2.8)</td>
<td>9 (6)</td>
<td>115 (17)</td>
<td>32 (9)</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Error rates are % errors out of errors and correct responses; in parentheses are standard errors of the mean computed from the analysis by participants and items, respectively. For response counts, dysfluency counts are shown in parentheses. Misc = Miscellaneous.
Table 4: Experiment 2 Semantic Integration and Plausibility Ratings

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun</th>
<th>Semantic Integration</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of</td>
<td>Singular</td>
<td>4.66 (.69)</td>
<td>5.61 (.84)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>4.35 (.94)</td>
<td>5.50 (.96)</td>
</tr>
<tr>
<td>With</td>
<td>Singular</td>
<td>4.42 (.92)</td>
<td>5.63 (.73)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>4.12 (.76)</td>
<td>5.50 (.81)</td>
</tr>
</tbody>
</table>

*Note.* Rating scales were 1 (not linked/implausible) to 7 (tightly linked/plausible). Standard deviations are shown in parentheses.
### Table 5: Experiment 2 Error Rates (%) and Response Counts by Condition

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun Number</th>
<th>Error Rate</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of</td>
<td>Singular</td>
<td>0 (0.0, 0.0)</td>
<td>0 (0)</td>
<td>67 (0)</td>
<td>148 (1)</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>19 (4.6, 5.2)</td>
<td>16 (0)</td>
<td>67 (1)</td>
<td>140 (1)</td>
<td>25</td>
</tr>
<tr>
<td>With</td>
<td>Singular</td>
<td>3 (1.1, 5.0)</td>
<td>2 (0)</td>
<td>62 (1)</td>
<td>147 (1)</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>10 (3.4, 3.7)</td>
<td>5 (0)</td>
<td>44 (2)</td>
<td>146 (5)</td>
<td>36</td>
</tr>
</tbody>
</table>

*Note.* Error rates are % errors out of errors and correct responses; in parentheses are standard errors of the mean computed from the analysis by participants and items, respectively. For response counts, dysfluency counts are shown in parentheses. Misc = Miscellaneous.
Table 6: Experiment 3 Semantic Integration and Plausibility Ratings

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun Number</th>
<th>Semantic Integration</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>For</td>
<td>Singular</td>
<td>4.65 (.68)</td>
<td>5.87 (.69)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>4.56 (.79)</td>
<td>5.86 (.80)</td>
</tr>
<tr>
<td>With</td>
<td>Singular</td>
<td>4.42 (.92)</td>
<td>5.63 (.73)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>4.12 (.76)</td>
<td>5.50 (.81)</td>
</tr>
</tbody>
</table>

*Note.* Rating scales were 1 (not linked/plausible) to 7 (tightly linked/plausible). Standard deviations are shown in parentheses. Data for *with* are repeated from Table 4.
Table 7: Experiment 3 Error Rates (%) and Response Counts by Condition

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun Number</th>
<th>Error Rate</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>For</td>
<td>Singular</td>
<td>0 (0.0, 0.0)</td>
<td>0 (0)</td>
<td>58 (6)</td>
<td>177 (26)</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>27 (4.1, 6.4)</td>
<td>13 (1)</td>
<td>35 (17)</td>
<td>161 (24)</td>
<td>36</td>
</tr>
<tr>
<td>With</td>
<td>Singular</td>
<td>3 (1.7, 5.0)</td>
<td>1 (4)</td>
<td>31 (10)</td>
<td>173 (34)</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>15 (2.9, 4.3)</td>
<td>5 (1)</td>
<td>29 (12)</td>
<td>164 (25)</td>
<td>39</td>
</tr>
</tbody>
</table>

Note. Error rates are % errors out of errors and correct responses; in parentheses are standard errors of the mean computed from the analysis by participants and items, respectively. For response counts, dysfluency counts are shown in parentheses. Misc = Miscellaneous.
Table 8: Experiment 4 Semantic Integration and Plausibility Ratings

<table>
<thead>
<tr>
<th>Semantic Role</th>
<th>Local Noun Number</th>
<th>Semantic Integration</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Singular</td>
<td>5.44 (.80)</td>
<td>5.70 (.70)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>5.53 (.63)</td>
<td>5.69 (.83)</td>
</tr>
<tr>
<td>Accompaniment</td>
<td>Singular</td>
<td>3.25 (.89)</td>
<td>4.35 (1.15)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>3.28 (.98)</td>
<td>4.03 (1.01)</td>
</tr>
<tr>
<td>Locative</td>
<td>Singular</td>
<td>4.09 (.79)</td>
<td>5.85 (.72)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>3.96 (.86)</td>
<td>5.71 (.64)</td>
</tr>
</tbody>
</table>

*Note.* Rating scales were 1 (not linked/implausible) to 7 (tightly linked/plausible). Standard deviations are shown in parentheses.
Table 9: Experiment 4 Error Rates (%) and Response Counts by Condition

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Local Noun Number</th>
<th>Error Rate</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Singular</td>
<td>2 (0.9, 2.1)</td>
<td>2 (0)</td>
<td>130 (9)</td>
<td>99 (9)</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>28 (5.2, 6.0)</td>
<td>34 (1)</td>
<td>89 (28)</td>
<td>87 (11)</td>
<td>30</td>
</tr>
<tr>
<td>Accompaniment</td>
<td>Singular</td>
<td>8 (2.1, 2.2)</td>
<td>10 (14)</td>
<td>116 (15)</td>
<td>68 (11)</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>15 (4.0, 2.9)</td>
<td>22 (9)</td>
<td>125 (17)</td>
<td>52 (9)</td>
<td>27</td>
</tr>
<tr>
<td>Locative</td>
<td>Singular</td>
<td>2 (2.1, 0.7)</td>
<td>2 (0)</td>
<td>123 (15)</td>
<td>211 (16)</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>12 (3.8, 3.5)</td>
<td>13 (6)</td>
<td>99 (20)</td>
<td>168 (26)</td>
<td>125</td>
</tr>
</tbody>
</table>

*Note.* Error rates are % errors out of errors and correct responses; in parentheses are standard errors of the mean computed from the analysis by participants and items, respectively. For response counts, dysfluency counts are shown in parentheses. Misc = Miscellaneous.
Table 10: Experiment 5 Semantic Integration and Plausibility Ratings

<table>
<thead>
<tr>
<th>Structure</th>
<th>Local Noun Number</th>
<th>Semantic Integration</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of-PP</td>
<td>Singular</td>
<td>3.52 (.79)</td>
<td>5.27 (.58)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>3.47 (.94)</td>
<td>4.99 (.65)</td>
</tr>
<tr>
<td>RC</td>
<td>Singular</td>
<td>3.57 (.87)</td>
<td>5.25 (.54)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>3.51 (.99)</td>
<td>5.24 (.58)</td>
</tr>
<tr>
<td>SC</td>
<td>Singular</td>
<td>2.94 (.72)</td>
<td>4.75 (.70)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>2.93 (.78)</td>
<td>4.67 (.83)</td>
</tr>
</tbody>
</table>

*Note.* Rating scales were 1 (not linked/plausible) to 7 (tightly linked/plausible). Standard deviations are shown in parentheses.
Table 11: Experiment 5 Error Rates (%) and Response Counts by Condition

<table>
<thead>
<tr>
<th>Structure</th>
<th>Local Noun</th>
<th>Error Rate</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Error</td>
</tr>
<tr>
<td>Of-PP</td>
<td>Singular</td>
<td>1 (0.6, 0.5)</td>
<td>1 (0)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>20 (3.6, 3.8)</td>
<td>32 (6)</td>
</tr>
<tr>
<td>RC</td>
<td>Singular</td>
<td>1 (0.4, 0.9)</td>
<td>2 (0)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>10 (2.8, 2.3)</td>
<td>21 (1)</td>
</tr>
<tr>
<td>SC</td>
<td>Singular</td>
<td>3 (1.2, 1.1)</td>
<td>4 (1)</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>1 (0.6, 0.6)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

Note. Error rates are % errors out of errors and correct responses; in parentheses are standard errors of the mean computed from the analysis by participants and items, respectively. For response counts, dysfluency counts are shown in parentheses. Misc = Miscellaneous.
Table 12: Stimuli Rated from Prior Studies

<table>
<thead>
<tr>
<th>Survey</th>
<th>Study</th>
<th>Exp</th>
<th>N</th>
<th>Factor(s) manipulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>Bock &amp; Miller (1991)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>16</td>
<td>distributivity and length; PP versions only</td>
</tr>
<tr>
<td>1, 2</td>
<td>Bock &amp; Miller (1991)&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>2</td>
<td>32</td>
<td>animacy of head noun (local noun always mismatched head in animacy) and relative concreteness of head and local noun</td>
</tr>
<tr>
<td>1, 2</td>
<td>Bock &amp; Cutting (1992)</td>
<td>1</td>
<td>32</td>
<td>length and construction (PP vs. RC)</td>
</tr>
<tr>
<td>1, 2</td>
<td>Bock &amp; Cutting (1992)</td>
<td>3</td>
<td>32</td>
<td>length and construction (PP vs. SC); only singular-head versions in both surveys</td>
</tr>
<tr>
<td>1, 2</td>
<td>Bock, Nicol, &amp; Cutting (1999)</td>
<td>1</td>
<td>36</td>
<td>head noun semantic number (singular and collective in Survey 1, plural in Survey 2)</td>
</tr>
<tr>
<td>1</td>
<td>Eberhard (1999)</td>
<td>2</td>
<td>12</td>
<td>distributivity; 4 additional items were from Bock &amp; Miller (1991, Experiment 1)</td>
</tr>
<tr>
<td>2</td>
<td>Bock &amp; Eberhard (1993)</td>
<td>3</td>
<td>8</td>
<td>head noun number and local noun plural morphology (regular vs. irregular)</td>
</tr>
</tbody>
</table>

Note. All rated stimuli had singular heads in Survey 1 and plural heads in Survey 2, except as noted. Ratings were always obtained for both the singular and plural local noun versions. All factors were manipulated within items except distributivity. Exp = Experiment. N = number of items.

<sup>a</sup>Error and correct agreement count data were provided by Kay Bock (p.c.). <sup>b</sup>Stimuli for rating were provided by Kay Bock (p.c.).
Table 13: Survey 1 and 2 Semantic Integration Ratings

<table>
<thead>
<tr>
<th>Study</th>
<th>Condition</th>
<th>Singular Head</th>
<th></th>
<th>Plural Head</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Match</td>
<td>Mismatch</td>
<td>Match</td>
<td>Mismatch</td>
</tr>
<tr>
<td>BM1</td>
<td>Long Distributive</td>
<td>4.94 (1.14)</td>
<td>5.18 (1.00)</td>
<td>4.68 (1.07)</td>
<td>4.61 (1.02)</td>
</tr>
<tr>
<td></td>
<td>Long Nondistributive</td>
<td>4.80 (.66)</td>
<td>4.40 (1.06)</td>
<td>4.09 (.93)</td>
<td>4.11 (.85)</td>
</tr>
<tr>
<td></td>
<td>Short Distributive</td>
<td>4.88 (.94)</td>
<td>4.58 (1.20)</td>
<td>4.86 (.96)</td>
<td>4.67 (1.23)</td>
</tr>
<tr>
<td></td>
<td>Short Nondistributive</td>
<td>4.96 (.49)</td>
<td>4.56 (.70)</td>
<td>4.89 (.89)</td>
<td>5.21 (.36)</td>
</tr>
<tr>
<td>BM2</td>
<td>Animete Head</td>
<td>4.14 (1.70)</td>
<td>3.93 (1.44)</td>
<td>4.32 (1.31)</td>
<td>4.10 (1.38)</td>
</tr>
<tr>
<td></td>
<td>Inanimate Head</td>
<td>3.54 (1.36)</td>
<td>3.61 (1.34)</td>
<td>3.85 (1.45)</td>
<td>4.17 (1.38)</td>
</tr>
<tr>
<td>BM2</td>
<td>More Concrete Head</td>
<td>3.97 (1.61)</td>
<td>3.77 (1.40)</td>
<td>3.97 (1.48)</td>
<td>4.15 (1.41)</td>
</tr>
<tr>
<td></td>
<td>Less Concrete Head</td>
<td>3.71 (1.51)</td>
<td>3.78 (1.40)</td>
<td>4.20 (1.31)</td>
<td>4.12 (1.35)</td>
</tr>
<tr>
<td>BC1</td>
<td>PP</td>
<td>4.81 (1.13)</td>
<td>4.77 (1.07)</td>
<td>5.25 (1.00)</td>
<td>5.11 (1.17)</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>4.73 (1.18)</td>
<td>4.77 (1.02)</td>
<td>4.99 (1.04)</td>
<td>5.12 (1.05)</td>
</tr>
<tr>
<td>BC3</td>
<td>Long PP</td>
<td>2.95 (.85)</td>
<td>2.94 (1.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long SC</td>
<td>2.62 (.80)</td>
<td>2.67 (.67)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short PP</td>
<td>3.01 (.95)</td>
<td>2.96 (.93)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Short SC</td>
<td>2.62 (.88)</td>
<td>2.62 (.76)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC</td>
<td>Singular/Plural Head</td>
<td>4.53 (1.11)</td>
<td>4.28 (1.15)</td>
<td>4.51 (1.35)</td>
<td>4.08 (1.31)</td>
</tr>
<tr>
<td></td>
<td>Collective Head</td>
<td>4.20 (1.23)</td>
<td>4.16 (1.26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>Distributive</td>
<td>5.56 (.50)</td>
<td>5.64 (.57)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nondistributive</td>
<td>5.17 (.91)</td>
<td>4.74 (1.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BE3</td>
<td>Regular Morphology</td>
<td>3.87 (1.21)</td>
<td>4.02 (1.19)</td>
<td>3.82 (1.16)</td>
<td>3.96 (1.18)</td>
</tr>
<tr>
<td></td>
<td>Irregular Morphology</td>
<td>4.27 (1.17)</td>
<td>4.36 (1.43)</td>
<td>3.98 (1.28)</td>
<td>4.12 (1.43)</td>
</tr>
</tbody>
</table>

*Note.* The rating scale was 1 (not linked) to 7 (tightly linked). Standard deviations are shown in parentheses. Studies are indicated by authors’ initials and experiment number (BM = Bock & Miller, 1991; BC = Bock & Cutting, 1992; BNC = Bock, Nicol, & Cutting, 1999; E = Eberhard, 1999; BE = Bock & Eberhard, 1993). Reported singular-head ratings are from Survey 1, and plural-head ratings are from Survey 2; except for Bock & Eberhard (1993), for which all ratings were collected in Survey 2.
Figure Captions

Figure 1: Subject noun phrase (NP) structural configurations for NPs containing argument prepositional phrases (PPs) (structure A or C) and non-argument PPs (structure A, B, or C). N = Noun, Det = Determiner.

Figure 2: Head-local mismatch error effect (plural local noun — singular local noun) versus plural local noun semantic integration rating (1-7; 7 = tightly linked) for Experiments 1-5. Each point represents one singular/plural local noun condition pair. Exp = Experiment.
