Journal of Experimental Psychology:
Learning, Memory, and Cognition
2013, Vol. 39, No. 2, 515–528

0278-7393/13/$12.00 DOI: 10.1037/a0029005

RESEARCH REPORT

Against Structural Constraints in Subject–Verb Agreement Production

Maureen Gillespie
University of Illinois Urbana–Champaign

Neal J. Pearlmutter
Northeastern University

Syntactic structure has been considered an integral component of agreement computation in language production. In agreement error studies, clause-boundedness (Bock & Cutting, 1992) and hierarchical feature-passing (Franck, Vigliocco, & Nicol, 2002) predict that local nouns within clausal modifiers should produce fewer errors than do those within phrasal modifiers due to structural differences; however, Gillespie and Pearlmutter (2011b) suggested structure might play a more limited role. Two studies examined whether the clause-boundedness effect would occur when prepositional phrase modifiers and relative clause modifiers were matched in properties likely to influence the timing of planning (Gillespie & Pearlmutter, 2011b; Solomon & Pearlmutter, 2004). In both studies, more errors occurred for plural local nouns, but the clause-boundedness effect was not observed. These findings suggest that agreement computation during production does not involve a hierarchical component.

Keywords: clause-boundedness, hierarchical feature-passing, sentence production, speech errors, subject–verb agreement

Theories of agreement production concerned with the mechanisms underlying the implementation of agreement have suggested that syntactic structure is particularly important for computing agreement relations. The clause-boundedness hypothesis suggests that agreement computation is only sensitive to information within the current clause (Bock & Cutting, 1992), and the hierarchical feature-passing hypothesis posits that agreement features are passed along the syntactic tree to their targets (Franck, Vigliocco, & Nicol, 2002). Alternatively, the scope of planning hypothesis (Gillespie & Pearlmutter, 2011b) explains agreement computation through processing that encodes the features of the agreement source and then retrieves them during the planning of the agreement target (see also Badecker & Kuminiak, 2007, for a related retrieval-based account); however, the studies supporting the scope of planning hypothesis did not test the influence of clausal structure on agreement nor the effects of hierarchical feature-passing over a limited amount of planned structure. The current study simultaneously tests these possibilities.

This article was published Online First June 25, 2012.

Maureen Gillespie, Department of Psychology, University of Illinois Urbana–Champaign; Neal J. Pearlmutter, Department of Psychology, Northeastern University.

Portions of this work were presented at the 2009 AMLaP Conference (Barcelona, Spain) and the 2010 CUNY Sentence Processing Conference (New York, NY). Maureen Gillespie was partially supported by NIH Grant 5T32HD055272 (primary investigator: Kay Bock). We thank Kay Bock for helpful feedback and for suggesting the comparison between tasks, and Mike Amato for creating the original versions of the Experiment 1 stimuli. In addition, we thank Libby Bernier, Jason Bran, Amy DiBattista, Shreya Divatia, Katelyn Flick, Ranya Gebara, Laura Goodman, Keith Levin, Lauren Mangold, Ian More, Carolyn Schulz, and Mariah Warren for help in collecting data and carefully transcribing and coding responses.

Correspondence concerning this article should be addressed to Maureen Gillespie, 603 East Daniel Street, Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL 61820. E-mail: mgillesp@illinois.edu

The first finding suggesting that hierarchical structure was a component of agreement production was the clause-boundedness effect. Bock and Cutting's (1992) Experiment 1 compared prepositional phrase (PP) modifier preambles (1a) to corresponding length-matched relative clause (RC) modifier preambles (1b), using an agreement error elicitation task in which participants recited each preamble aloud and completed it as a sentence. Subject–verb agreement error rates were larger when the local noun was plural (e.g., *books*) than when the local noun was singular (e.g., *book*), the standard "mismatch effect" (Bock & Miller, 1991), but this difference was larger for PP than for RC cases (see also Solomon & Pearlmutter, 2004, Experiment 5). Bock and Cutting (1992) suggested that this was because clauses are planned independently, so elements within separate clauses are less likely to interfere with each other than are elements within the same clause.

1a. The editor of the history book(s) (PP)

1b. The editor who rejected the book(s) (RC)

An alternative explanation for the clause-boundedness effect comes from the hierarchical feature-passing hypothesis, which provides a structure-based mechanism for implementing agreement (Eberhard, Cutting, & Bock, 2005; Franck et al., 2002; Hartsuiker, Antón-Méndez, & van Zee, 2001; Vigliocco & Hartsuiker, 2002).[1] On this view, agreement is computed using the syntactic tree structure of a sentence, with number features being passed up through the subject noun phrase (NP) and then to the

[1] More recent work by Franck and colleagues (Franck et al., 2006, 2010, 2008) does not rely on hierarchical feature-passing as a mechanism for implementing agreement, but it does rely on the hierarchical depth of the local noun within the subject NP as an explanatory factor, and it makes the same predictions as hierarchical feature-passing for the critical stimuli in the current studies. We consider this work in more detail in the General Discussion.

verb phrase. Mismatch effects occur when a plural feature is inadvertently passed too far up the tree, overwriting the number from the head noun with the number from a local noun. Franck et al. (2002) provided the most direct test of hierarchical feature-passing, using subject NP preambles containing two PPs, as in 2. Their stimuli had a descending hierarchical structure in which each PP modified the immediately preceding noun, and the local nouns (*flight* and *canyon* in 2) varied in number. The hierarchical feature-passing hypothesis predicts more errors for preambles like 2b than for preambles like 2c because the second noun (N2; *flight[s]*) is hierarchically closer to the verb than N3 (*canyon[s]*) is, and fewer feature-passing errors would have to occur for N2's plural to interfere with agreement than for N3's plural to interfere. Franck et al. found that the N2 mismatch effect was larger than the N3 mismatch effect in both English and French and argued for a hierarchical feature-passing account of subject–verb agreement over an account in which interference increases with linear proximity to the verb. Hierarchical feature-passing explains the clause-boundedness effect because local nouns in PPs are hierarchically closer to the verb than are local nouns in RCs by virtue of the additional clause-internal structure needed for RCs but not PPs (Franck et al., 2002; see Solomon & Pearlmutter, 2004, for discussion).

2a. The helicopter for the flight over the canyon

2b. The helicopter for the flights over the canyon

2c. The helicopter for the flight over the canyons

2d. The helicopter for the flights over the canyons

Franck et al. (2002) provided an alternative to clause-boundedness as the explanation for Bock and Cutting's (1992) results, but their data cannot rule out clause-boundedness as one of multiple factors influencing agreement. Two later experiments in French by Franck and colleagues might be considered as stronger tests, however: Franck and Nicol (in preparation, as reported in Franck et al., 2004) compared PP modifier cases like 3a to "clausal adjunct" cases like 3b (each with a singular local noun control) and found higher mismatch error rates for the clausal cases. This conflicts with the prediction of clause-boundedness, which, as in the case of Bock and Cutting, is that the mismatch effect should be larger for the PPs than for the clauses. Interpreting this result is difficult, however: First, the modifier manipulation was between items, and the PP and clausal stimuli were quite different in content, aside from being length- and animacy-matched. Second, the syntactic, semantic, and discourse properties of clausal adjuncts like 3b are relatively unstudied. Both concerns make it hard to determine what is responsible for the difference between the conditions, and Franck et al. (2004) in fact proposed that the clausal cases are handled by a separate process from those responsible for the phrasal modifier case.

3a. *La gagnante des derniers championnats* (The winner of the last championships)

3b. *La grand-mère, en parlant aux filles* (The grandmother, while talking to the girls)

Franck, Soare, Frauenfelder, and Rizzi (2010, Experiment 5) also compared two structural conditions in only one of which the local noun came from a separate clause, as in 4 (each had a singular local noun control), in which the uppercase verb was given in advance to the participants in infinitival form, with no number marking. Rather than a sentence beginning, participants saw the entire sentence except the critical verb, with the position to be filled by the verb indicated by an underscore. In 4a, the local noun *traîtres* (*traitors*) is understood as the object of *jugera* (*judge*), and that position is part of a separate clause from *victime* (*victim*), the head noun. Thus, interference from *traîtres* should be weaker in this case than in 4b, in which *traîtres* is understood as the object of the critical verb *défend* (*defends*), part of the same clause as *victime.*

4a. *Voilà les traîtres que la victime DIT qu'on jugera* (Here are the traitors that the victim SAYS that we will judge).

4b. *Voilà les traîtres que la victime DEFEND malgré sa douleur* (Here are the traitors that the victim DEFENDS despite his illness).

Franck et al. (2010) in fact found no difference in mismatch effects between the two structures, suggesting that the clause boundary in 4a did not reduce the likelihood of interference from the local noun. This result argues against a version of clause-boundedness that attends to some version of an element's base-generated position in a structure (roughly, the position where the element would be interpreted for the purpose of computing meaning; e.g., for *traîtres* in 4a, immediately after *jugera*), but clause-boundedness can alternatively be formulated over positions in the surface string or with sensitivity to both base-generated and surface positions; in these cases, it would likely not have a basis for distinguishing the two structures in 4, predicting the lack of a difference in mismatch effects that Franck et al. (2010) found. One other issue with interpreting this result arises from the nature of the task (and the stimuli): Participants were shown the verb to use, as well as the entire sentence to utter (with the verb-insertion position marked); they only had to inflect the verb, insert it, and recite the combination. Furthermore, the only difference between the structure conditions was the material following the verb-insertion position. The combination of these properties might have made participants less sensitive to the difference in structure and created something more like a two-alternative forced-choice task than a typical sentence production task (see Gillespie & Pearlmutter, 2011b, for related discussion). Thus while both Franck et al. (2004) and Franck et al. (2010) provided some suggestive evidence about clause-boundedness, whether or not it plays any role in agreement production remains to be established.

In addition to structural properties, semantic and temporal properties that influence the timing of planning also seem to affect agreement computation. Solomon and Pearlmutter (2004) hypothesized that semantic integration (i.e., the degree to which elements within a phrase are linked at the message level) affects the timing of planning of elements within a phrase, such that elements of more semantically integrated phrases are more likely to be planned overlappingly. Solomon and Pearlmutter manipulated local noun number in NP PP stimuli and compared integrated cases (e.g., *The*

pizza with the yummy topping[s]) to corresponding unintegrated ones (e.g., *The pizza with the tasty beverage[s]*). Across a series of experiments, they found larger mismatch effects for integrated than for unintegrated conditions, supporting the hypothesis that overlap in planning leads to increased interference during agreement computation (for evidence from exchange errors, see DiBattista & Pearlmutter, 2011; Pearlmutter & Solomon, 2007).

Gillespie and Pearlmutter (2011b) noted that Franck et al.'s (2002) stimuli had a semantic integration confound: The head noun (N1; *helicopter* in 2) and N2 were more semantically integrated were than N1 and N3, so semantic integration might explain Franck et al.'s (2002) results. In addition, Franck et al. (2002) did not discuss the possibility that a local noun's linear proximity to N1 might increase error rates, which could also explain the results they attributed to hierarchical distance. Gillespie and Pearlmutter's (2011b) Experiment 1 used NP PP PP preambles that varied structure, such that half the preambles had a descending structure like Franck et al.'s (2002) preambles, and the other half had a flat structure with both PPs modifying N1. Critically, semantic integration of the N1–N2 pairs was equated across structures, as was the semantic integration of the N1–N3 pairs. Gillespie and Pearlmutter's (2011b) found no effect of structure on the size of mismatch effects; instead, only linear proximity to N1 affected error rates: N2 plurals elicited larger mismatch effects than did N3 plurals. Gillespie and Pearlmutter's (2011b) Experiment 2 used NP PP PP preambles with a flat structure and manipulated semantic integration and linear distance, and it showed a combination of linear distance and semantic integration effects. Gillespie and Pearlmutter (2011b) proposed a scope of planning account of agreement production, predicting more agreement errors when a plural local noun is planned within the scope of (i.e., close in time to) a singular head noun, with semantic integration and linear order combining to influence planning time, independent of hierarchical distance (see Badecker & Kuminiak, 2007, and Nicol, 1995, for details of other proposals that consider planning time as a factor affecting agreement computation).

This scope of planning account can explain Franck et al.'s (2002) results and many other effects reported in the agreement literature without a need for hierarchical feature-passing, which raises the question of whether agreement computations are constrained directly by structure at all. But while Gillespie and Pearlmutter (2011b) argued against a hierarchical account of existing agreement data and suggested an alternative mechanism, they could not rule out two possibilities: that hierarchical feature-passing is the mechanism underlying all agreement computation but that its effects are constrained by scope of planning (errant feature-passing cannot occur from within as-yet-unplanned constituents) or that feature-passing applies only to or around clause boundaries. Clause-boundedness itself, the other main proposed structural constraint (but cf. Franck et al., 2010), also cannot be explained by scope of planning: Bock and Cutting's (1992) PP and RC stimuli were matched for length in syllables (linear distance from head to local noun), and Solomon and Pearlmutter (2004) showed they were also matched for semantic integration; Solomon and Pearlmutter also replicated the clause-boundededness effect with their own set of integration- and length-matched stimuli. Thus, either or both of hierarchical feature-passing or clause-boundedness might at least constrain agreement computation.

The current studies investigated this question by reexamining the clause-boundedness effect. While PPs and RCs in previous studies were matched on length and semantic integration, they differed in at least two other potentially relevant ways: First, the RCs linked the head and local noun with a content word (a semantically rich verb), whereas the PPs used a function word (the preposition). Second, the PPs and RCs differed in overall meaning; and various conceptual properties have been shown to influence agreement error rates, either directly (e.g., distributivity, noun conceptual number; Eberhard et al., 2005) or indirectly (e.g., concreteness, Eberhard, 1999). Experiment 1 examined whether the clause-boundedness effect was observed when PPs and RCs were matched in overall meaning and used function words to link the head and local noun, and Experiment 2 examined whether the presence of content verbs in RC conditions contributed to the clause-boundedness effect observed in previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004).

Experiment 1

The goal of Experiment 1 was to test the predictions of clause-boundedness and hierarchical feature-passing while controlling for semantic integration, linear distance, and the two properties discussed above. The only difference between the PPs and RCs used in this experiment was that the PPs contained the preposition *with,* in its attribute/possessive sense, while the RCs contained the verb *had,* in its relatively, semantically light possessive sense. RCs always contained the complementizer *that,* making RCs exactly one word longer than PPs. Thus, the PPs and RCs were matched in number of adjectives, properties of the linking word, and general meaning; however, they differed in the clausal structure and the local noun's hierarchical distance to the subject NP node (see General Discussion and Solomon & Pearlmutter, 2004, Experiment 5). If the difference in error rates between PPs and RCs in previous studies was due to structure instead of any of the other factors that varied, the PP mismatch effect here should be greater than the RC mismatch effect.

A secondary goal was to compare the two commonly used versions of the agreement elicitation task: (a) recall tasks, which require speakers to listen to or read preambles, hold them in memory, then repeat them to complete a full sentence and (b) no-recall tasks, which require speakers to read preambles aloud and then complete them as full sentences. Both tasks have shown structural effects (Bock & Cutting, 1992, used an auditory-presentation recall task; while Solomon & Pearlmutter, 2004, found essentially identical results using a visual-presentation no-recall task), but the timing of the planning may nevertheless be different across them due to differing memory demands or differences in the influence of comprehension processes during production. Task was manipulated between-participants to examine these possibilities.

Method

Participants. Fifty-nine Northeastern University undergraduates participated in the no-recall task, but one participant was excluded for being unable to read the preambles before they disappeared. Sixty-four Northeastern University students participated in the recall task, but two participants were excluded because

they began speaking before the signal tone on nearly every trial. All participants were native English speakers, and they received course credit for their participation; no participant provided data for more than one part of the experiment.

Materials and design. Twenty-four stimulus sets like that shown in Table 1 were constructed. Each began with a head NP (e.g., *The pizza*) followed by a modifier containing a local noun (e.g., *slice[s]*). The head noun was always singular, and the four different versions of an item were created by varying modifier type and (local) noun number. The modifier was either a PP or an RC and was a description of an attribute of the head noun. PP modifiers began with the preposition *with,* and they were followed by a local NP consisting of a determiner, adjective, and noun. RC modifiers began with the complementizer *that* and the verb *had,* followed by the same local NP. As a result, the RCs were always exactly one syllable or word longer than the corresponding PPs.

In addition to the critical items, 88 fillers were included. Twenty-four of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 14–15 participants in the no-recall task and by 15–16 participants in the recall task. The complete list of critical stimuli is shown in Appendix A.

Stimulus norming. The 24 critical stimuli were normed for semantic integration by 51 participants (two more were excluded for failing to follow instructions). The four different versions of each of the 24 items, along with 24 fillers intended to cover the full rating scale, were rated using a 1 (*loosely linked*) to 7 (*tightly linked*) scale, following the procedure described in Solomon and Pearlmutter (2004). The four versions of each item were counterbalanced across four lists such that exactly one version of each stimulus item appeared in each list, and 12–14 ratings were obtained for each version. Table 1 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression (Baayen, Davidson, & Bates, 2008) on these data (random factors: participant, item; fixed effects: local noun number, modifier, and their interaction) revealed no main effects nor an interaction (all $|t|$s < 1.2, ps $> .23$).[2]

Apparatus and procedure. Each participant was run individually in the main experiment using either the no-recall or recall task. In the no-recall task, participants read each visually presented preamble aloud as soon as it appeared and added an ending that formed a complete sentence. In the recall task, participants read each visually presented preamble silently as soon as it appeared and then, after a tone, repeated the preamble aloud and added an ending that formed a complete sentence. Participants were not instructed as to how they should formulate a completion, only that they should form a complete sentence.

In both tasks, on each trial, a fixation cross appeared at the left edge of the display for 1,000 ms, followed by the preamble. Each preamble was presented for the longer of 1,000 ms or 50 ms/character. After the preamble disappeared, the screen was blank for 2,000 ms, followed by a prompt to begin the next trial. In the recall task, a tone was presented immediately after the preamble disappeared to indicate that participants could begin speaking. A PC running the MicroExperimental Laboratory software package (Schneider, 1988) controlled stimulus presentation, and participants' responses were recorded to CD for analysis, using a Shure SM58 microphone connected to a Mackie 1202-VLZ Pro mixer/preamp and an Alesis Masterlink ML-9600 (OS v2.20) CD recorder. Five practice items preceded the 112 trials.

Scoring. All responses were transcribed and assigned to one of four coding categories: (a) correct, if the participant repeated the preamble correctly exactly once, produced an inflected verb immediately after the preamble, and used a verb form that was correctly marked for number; (b) error, if all the criteria for correct responses were met, but the verb form failed to agree in number with the subject; (c) uninflected, if all the criteria for correct responses were met, but the verb was uninflected; and (d) miscellaneous, if the participant made an error repeating the preamble, if a verb did not immediately follow the preamble, if participants began speaking before the tone in the recall task, or if the response did not fall into any of the other categories. Trials in which a participant made no response were excluded from all analyses. If the participant produced a dysfluency (e.g., pauses, coughs) during or immediately after producing the preamble and went on to produce a correct, error, or uninflected response, the scoring category and the dysfluency were recorded. On miscellaneous trials, dysfluencies were not separately counted.

Results

Table 2 shows the counts of each response type by task, modifier, and noun number, with the number of responses containing a dysfluency in parentheses. Separate analyses were performed for error rates (the proportion of error responses out of error plus correct responses), uninflected rates (the proportion of uninflected responses out of total scorable responses), and miscellaneous rates (the proportion of miscellaneous responses out of total scorable responses). The reported error and uninflected analyses included dysfluencies, and unless otherwise noted, the patterns were identical if dysfluency cases were excluded.

Performing analyses of variance (ANOVAs) on proportion data is problematic and may produce spurious results; Jaeger (2008) instead suggested analyzing such data using logit mixed-effect models. However, the error rates produced in subject–verb agreement studies are often extremely low, creating problems in applying the logit link function during model fitting (the log odds of proportions near 0 approach negative infinity). Thus, following Barr (2008), the data were analyzed using empirical logit weighted linear regression, aggregating separately over participants and items. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed, with noun number, modifier, task, and all interactions as sum-coded fixed effects (t tests of parameter estimates are identified as t_1 for the by-participant analysis and as t_2 for the by-item analysis). We also computed corresponding ANOVAs on arcsine-transformed proportions (Cohen & Cohen, 1983), including all 58 participants from the no-recall task and 56 of the 62 participants

[2] All regression analyses were performed in R (R Development Core Team, 2010) with the languageR package (Baayen, 2008). Models were fit using the lme4 package (Version 0.999375-37), and p values were obtained using the MCMC sampling function in the coda package (Version 0.14-2).

Table 1
Experiment 1 Stimuli and Semantic Integration Ratings by Condition

Modifier	Noun number	Example	Semantic integration
PP	SP	The pizza with the missing slices	5.56 (1.22)
	SS	The pizza with the missing slice	5.65 (1.20)
RC	SP	The pizza that had the missing slices	5.58 (1.28)
	SS	The pizza that had the missing slice	5.53 (1.35)

Note. The semantic integration rating scale was 1 (*loosely linked*) to 7 (*tightly linked*); standard deviations are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local noun.

from the recall task (six were excluded because they were missing data in one or more cells). Results from ANOVA analyses are only reported when they differed from the regression analyses.

Agreement errors. Figure 1 shows untransformed error rates by condition collapsed over task, and Table 3 shows the weighted linear regression effect estimates. Errors were more likely when the local noun was plural than when it was singular, but there was no main effect of modifier and, critically, no interaction of noun number and modifier.

The tasks did not differ in the main analyses, but when dysfluencies were excluded, there were more errors in the no-recall task than in the recall task (significant by participants, marginal by items). The interaction of task and modifier was marginal by participants and by items, such that in the no-recall task RCs yielded more errors than PPs, but in the recall task, PPs yielded more errors than did RCs; however, separate analyses on each task showed no main effects of modifier. The interaction of task and modifier reached significance in the ANOVA by participants when dysfluencies were excluded. The interaction of task and noun number only reached significance in the ANOVA by items when dysfluencies were excluded, indicating that the mismatch effect was larger in the no-recall task than in the recall task. There was no hint of a three-way interaction in any analysis; we nevertheless also examined the Noun Number × Modifier interaction for each task separately, and neither was reliable (all $|t|$s < 1, ps $> .27$).

Uninflected rates. The interaction of task and modifier was significant by items ($t_1 = 1.45$, $p = .21$; $t_2 = 2.14$, $p < .05$), with higher uninflected rates for RCs than PPs in the recall task and higher rates for PPs than RCs in the no-recall task. The interaction of local noun number and modifier was marginal by items ($t_1 = -1.66$, $p = .14$; $t_2 = -3.18$), with RCs yielding higher uninflected rates for singular than for plural local noun cases and PPs yielding nearly equal uninflected rates for the two; this effect was nonsignificant when dysfluencies were excluded. Also, when dysfluencies were excluded, uninflected responses were more likely for singular local nouns ($t_1 = -1.70$, $p = .09$; $t_2 = -3.67$, $p < .05$). No other main effects or interactions approached significance (all $|t|$s < 2.6, ps $> .11$).

Miscellaneous rates. Miscellaneous responses were more likely for plural than singular local nouns ($t_1 = 3.74$, $t_2 = 4.06$, ps $< .01$), for RCs than for PPs ($t_1 = 2.94$, $t_2 = 3.54$, ps $< .05$), and in the recall task than in the no-recall task ($t_1 = 3.15$, $t_2 = 6.02$, ps $< .001$). There were no interactions (all $|t|$s < 1.9, ps $> .11$).

Discussion

The large noun number effect replicates essentially all studies examining mismatch effects: With singular heads, agreement error rates are larger when the local noun is plural than when it is singular (e.g., Bock & Miller, 1991; Eberhard, 1997). However, unlike similar previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004), modifier and noun number did not interact, indicating equal mismatch effects for PPs and RCs. Thus, the current study provided no evidence for structural effects on agreement when other differences between PPs and RCs were minimized. Because this experiment directly tested predictions of clause-boundedness and hierarchical feature-passing by manipulating clausal structure, thus varying the number of syntactic nodes

Table 2
Experiment 1 Response Counts by Task and Condition

Task	Modifier	Noun number	Error	Correct	Uninflected	Misc	No resp
No recall	PP	SP	18 (4)	196 (42)	95 (29)	37	2
		SS	0 (0)	230 (50)	84 (20)	31	3
	RC	SP	21 (1)	210 (49)	66 (16)	50	1
		SS	1 (0)	215 (53)	88 (23)	43	1
Recall	PP	SP	18 (3)	195 (11)	91 (12)	67	1
		SS	2 (0)	222 (27)	98 (9)	46	4
	RC	SP	8 (3)	186 (21)	90 (15)	86	2
		SS	0 (0)	207 (24)	107 (14)	57	1
Total			68 (11)	1,661 (277)	719 (138)	417	15

Note. Dysfluency counts are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local; Misc = miscellaneous; No resp = no response.

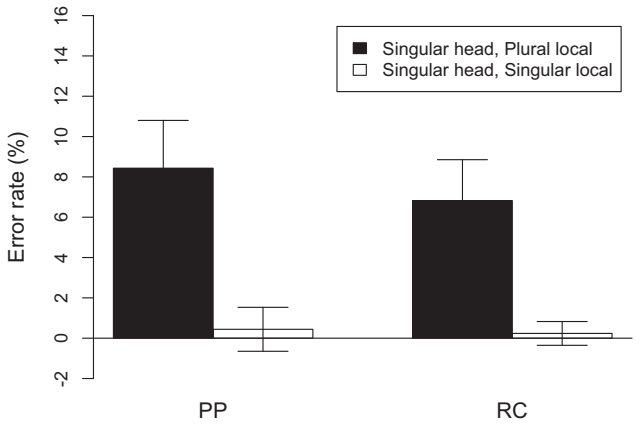

Figure 1. Experiment 1 untransformed agreement error rates as a function of modifier and noun number. Error bars indicate ±1 standard error of the mean, computed by items. PP = prepositional phrase; RC = relative clause.

a local noun's plural feature would have to pass through in order to influence agreement computation, these findings argue for an account of agreement production that does not involve a hierarchical component.

This result goes beyond Gillespie and Pearlmutter's (2011b), which had left open the possibility that hierarchical feature-passing could be the mechanism underlying agreement production, as long as the set of feature sources for feature-passing was constrained by the scope of planning. This possibility would allow a hierarchical feature-passing theory a second explanation for clause-boundedness effects (in addition to hierarchical distance differences), if local nouns in RCs were less likely than those in PPs to be planned overlappingly with the head. The presence of equal interference from local nouns in PPs and in RCs in the current study indicates that the relevant scope of planning did not vary across modifier type, ruling out a hierarchical explanation in these terms as well. Thus, this finding is incompatible with even a highly constrained use of feature-passing in agreement production.

A secondary goal of this study was to determine whether task affected error rates, but there was no three-way interaction of task, modifier, and noun number. Because the recall task requires speakers to hold the preamble in memory prior to repeating it and completing a sentence, interference could have arisen during retrieval of the preamble. This would have led to more agreement errors overall in the recall task if number information was susceptible to this interference. The recall task did increase miscellaneous errors, but agreement error rates tended to be higher in the no-recall task than in the recall task, suggesting that recall was not responsible for agreement error production. Both tasks also involve comprehension, which could be another source of interference (for discussion, see Gillespie & Pearlmutter, 2011b). While the comprehension component differed between tasks (concurrent with production in the no-recall task, prior to production in the recall task), mismatch effects were equal. Overall, the current findings indicate that the two tasks produce very similar error patterns and suggest that retrieval and comprehension processes cannot be entirely responsible for number interference effects observed in error elicitation paradigms.

While there were no significant interactions with task, and neither task separately showed a reliable interaction between noun number and modifier, the mismatch effect pattern in the recall task was numerically in the direction predicted by structural accounts (a larger mismatch effect for PPs than for RCs; see Table 2), whereas the pattern was (numerically) in the opposite direction for the no-recall task. Given that the argument against structural accounts depends on there being no interaction, Experiment 2 was conducted to gather as much additional data as possible in the task that came closest to showing an interaction in the direction predicted by structural accounts, while additionally testing a possible explanation for why the clause boundedness effect was eliminated in Experiment 1.

Experiment 2

While Experiment 1 showed no evidence of mismatch effects being affected by hierarchical structure, it also did not provide an explanation for the structural effects observed in earlier studies that manipulated clausal structure (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004). One difference between Experiment 1 and those studies is that the RC modifiers in Experiment 1 did not contain semantically rich content verbs, whereas the RC modifiers

Table 3
Experiment 1 Agreement Error Rate Results (Weighted Empirical Logit Linear Regression)

	By participants			By items		
Effect	β	*SE*	t_1	β	*SE*	t_2
Noun number (SP)	.53	.08	7.03*	.98	.12	8.15*
Modifier (RC)	−.03	.08	−0.40	−.07	.12	−0.56
Noun Number × Modifier	−.13	.15	−0.90	−.16	.24	−0.65
Task (Recall)	.01	.13	0.04	−.03	.12	−0.27
Task × Noun Number	−.14	.15	−0.90	−.30	.24	−1.25
Task × Modifier	−.18	.15	−1.18†	−.38	.24	−1.58†
Task × Noun Number × Modifier	−.06	.30	−0.21	−.10	.48	−0.21

Note. The level shown in parentheses for each variable was sum-coded +0.5 and the other level −0.5, so betas (βs) estimate the difference between the two levels of the variable in log-odds space. SP = singular head, plural local noun; RC = relative clause.
† $p < .10$. * $p < .001$.

in Bock and Cutting (1992) and Solomon and Pearlmutter (2004) did. Content words appear to be processed differently from function words in language production across a range of tasks and measures (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Dell, 1990; Jescheniak & Levelt, 1994), and some of these differences in turn might lead to differences in the planning time of local nouns relative to head nouns, which would then influence the agreement computation process under Gillespie and Pearlmutter's (2011b) scope of planning hypothesis. Thus, Experiment 2 examined whether the mere presence of a content verb within an RC modifier would reduce the mismatch effect.

Experiment 2 used preambles like those in Experiment 1, manipulating local noun number and modifier type, but with three modifier conditions (see Table 4 for an example item): As in Experiment 1, PP modifiers contained the preposition *with,* in its attribute/possessive sense. The RC-light condition was the same as the RC condition in Experiment 1, using the verb *had* in its relatively semantically light possessive sense, whereas the RC-content condition replaced *had* with a content verb (e.g., *contained, included*) that created a similar attribute/possessive relationship between the head noun and the local noun. The two RC conditions always contained the complementizer *that,* making them exactly one word longer than corresponding PPs. Thus, as in Experiment 1, the PPs and the two RC conditions were matched in number of adjectives and general meaning, but the PP conditions differed from the two RC conditions in clausal structure and the local noun's hierarchical distance to the subject NP node. Experiment 1 showed that PPs and RCs matched in meaning and linking word properties yielded equivalent mismatch effects, so if the difference in mismatch effects between PPs and RCs in earlier studies resulted from the presence of a content linking word in RCs, compared with a function linking word in PPs, the PP and RC-light mismatch effects in Experiment 2 should be equal, replicating Experiment 1, and both should be larger than the RC-content mismatch effect.

Method

Participants. One hundred seventy-three Northeastern University undergraduates participated. One participant accidentally completed the experiment twice, so the data from the second run were excluded. Data from five participants were excluded because the participants were nonnative English speakers, and data from one participant were lost due to a recording failure, leaving 167 participants' data to be analyzed. All participants received course credit for their participation. No participant provided data for more than one part of the experiment, and no participant had provided data in Experiment 1.

Materials and design. Twenty-four stimulus sets like that shown in Table 4 were constructed; most were modified versions of the stimuli from Experiment 1. Each began with a head NP (e.g., *The pizza*) followed by a modifier containing a local noun (e.g., *topping[s]*). The head noun was always singular, and the six different versions of an item were created by varying modifier type and (local) noun number. The modifier was a PP, an RC containing a light verb (RC-light), or an RC containing a content verb (RC-content); all modifiers described an attribute of the head noun. PP modifiers began with the preposition *with* and were followed by a local NP consisting of a determiner, adjective, and noun. RC-light modifiers began with the complementizer *that* and the verb *had,* followed by the same local NP. RC-content modifiers were identical to RC-light modifiers, but *had* was replaced by one of five possible content verbs (*contained, displayed, featured, held, included*). As in Experiment 1, the RC versions were always exactly one word longer than the corresponding PPs. The RC-content conditions were on average 1.4 syllables longer than the RC-light conditions and 2.4 syllables longer than the PP conditions.

In addition to the critical items, 108 fillers were included. Twenty-four of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in six counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 26–30 participants. The complete list of critical stimuli is shown in Appendix B.

Stimulus norming. Using Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011), the 24 critical stimuli were normed for semantic integration by 100 participants, but data from 10 participants were excluded due to a recording failure. The six different versions of each of the 24 items, along with 24 fillers intended to cover the full rating scale, were rated using instructions like those in Experiment 1 but modified slightly for presentation on Mechanical Turk. The six versions of each item were counterbalanced across 90 lists created using the software in Gibson, Piantadosi, and Fedorenko (2011), and 15 ratings were obtained for each version. Table 4 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression (Baayen et al., 2008) on these data (random factors: participant and item intercepts;

Table 4
Experiment 2 Stimuli and Semantic Integration Ratings by Condition

Modifier	Noun number	Example	Semantic integration
PP	SP	The pizza with the yummy toppings	5.34 (1.63)
	SS	The pizza with the yummy topping	5.37 (1.50)
RC-light	SP	The pizza that had the yummy toppings	5.35 (1.55)
	SS	The pizza that had the yummy topping	5.23 (1.63)
RC-content	SP	The pizza that included the yummy toppings	5.51 (1.53)
	SS	The pizza that included the yummy topping	5.33 (1.59)

Note. The semantic integration rating scale was 1 (*loosely linked*) to 7 (*tightly linked*); standard deviations are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local noun.

fixed effects: local noun number, modifier type, and their interaction) was performed. Local noun number was entered into the model as a sum-coded predictor, and modifier type was entered into the model as a treatment-coded predictor with the PP condition serving as the baseline.

There were no effects of noun number or modifier type ($|t|$s < 1.14, ps $> .26$). The noun number effect was equivalent in the PP and RC-light conditions ($t = 1.26$, $p > .21$). However, the noun number effect was marginally larger in the RC-content condition when compared to the PP condition ($t = 1.78$, $p = .08$), with the RC-content singular head, plural local noun (SP) condition receiving higher integration ratings than its corresponding singular head, singular local noun (SS) condition and the PP condition showing nearly identical ratings for the SP and SS versions. A separate analysis on the SP conditions (the conditions most likely to produce agreement errors) revealed no integration difference between the PP and RC-light conditions ($t = 0.20$, $p > .84$) and revealed that the RC-content condition was rated as more integrated than the PP condition ($t = 2.02$, $p < .05$). Potential effects of this difference on error rates are addressed in the Discussion (see Footnote 3).

Apparatus and procedure. Each participant was run individually in the main experiment. The procedure was identical to the recall version of the task used in Experiment 1. Five practice items preceded the 132 trials.

Scoring. Scoring was identical to that in Experiment 1.

Results

Table 5 shows the counts of each response type by modifier and noun number, with the number of responses containing a dysfluency in parentheses. As in Experiment 1, separate analyses were performed for error rates, uninflected rates, and miscellaneous rates. The reported error and uninflected analyses included dysfluencies, but the patterns were identical if dysfluency cases were excluded.

Also as in Experiment 1, the data were analyzed using empirical logit weighted linear regression, aggregating separately over participants and items. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed with noun number, modifier, and their interaction as fixed effects. Noun number was included as a sum-coded predictor, and modifier was included as a treatment-coded predictor, with PP as the base level. While not as critical to the predictions, models including only the RC conditions were also constructed, as the main analyses do not provide a direct comparison of these cases. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed on the RC data with noun number, modifier, and their interaction as fixed effects.

Corresponding ANOVAs were conducted on arcsine-transformed proportions (Cohen & Cohen, 1983), including 135 participants (32 were excluded because they were missing data in one or more cells). Results from ANOVA analyses are only reported when they differed from the regression analyses.

Agreement errors. Figure 2 shows untransformed error rates by condition, and Table 6 shows the weighted linear regression effect estimates. Errors were more likely when the local noun was plural than when it was singular. However, errors were not more likely in the PP condition overall compared to either RC condition, and critically, the mismatch effect did not differ for PP cases compared with RC-light cases nor for PP cases compared with RC-content cases. The models analyzing the RC conditions alone indicated that more errors were produced when the local noun was plural than when it was singular ($t_1 = 5.45$, $t_2 = 9.14$, ps $< .001$), but there was no main effect of modifier ($|t|$s < 1, ps $> .28$), and the RC-light and RC-content mismatch effects were equal ($|t|$s < 1.13, ps $> .14$).

Uninflected rates. Singular local nouns yielded uninflected responses marginally more often (by participants only) than plural local nouns ($t_1 = -1.57$, $p = .09$; $t_2 = -2.94$, $p = .17$). In addition, PP conditions yielded uninflected responses marginally more often (by participants only) than RC-light conditions did ($t_1 = -1.63$, $p = .08$; $t_2 = -3.35$, $p = .14$), and reliably more often than the RC-content conditions did ($t_1 = -3.91$, $t_2 = -7.72$, ps $< .001$). There were no Noun Number $\times$ Modifier interactions ($|t|$s < 2.75, ps $> .18$). The models analyzing the RC conditions alone indicated that RC-content conditions yielded fewer uninflected responses than RC-light conditions ($t_1 = -2.28$, $t_2 = -4.52$, ps $< .05$), with no effect of noun number and no interaction ($|t|$s < 1, ps $> .49$).

Miscellaneous rates. Miscellaneous responses were more likely for plural nouns than for singular local nouns ($t_1 = 2.14$, $t_2 = 4.08$, ps $< .05$), for RC-light than for PP conditions ($t_1 = 4.20$, $t_2 = 7.86$, ps $< .001$), and for RC-content than for PP conditions ($t_1 = 4.94$, $t_2 = 8.70$, ps $< .001$). There were no interactions (all $|t|$s < 2.54, ps $> .18$). The models analyzing the RC conditions alone showed no main effects and no interaction (all

Table 5
Experiment 2 Response Counts by Condition

Modifier	Noun number	Error	Correct	Uninflected	Misc	No resp
PP	SP	33 (3)	392 (15)	138 (12)	96	9
	SS	1 (0)	433 (15)	165 (11)	68	1
RC-light	SP	36 (3)	357 (17)	133 (7)	140	2
	SS	2 (0)	409 (11)	128 (7)	124	5
RC-content	SP	27 (2)	388 (23)	103 (4)	147	3
	SS	2 (0)	416 (23)	114 (10)	135	1
Total		101 (8)	2395 (104)	781 (51)	710	21

Note. Dysfluency counts are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local; Misc = miscellaneous; No resp = no response.

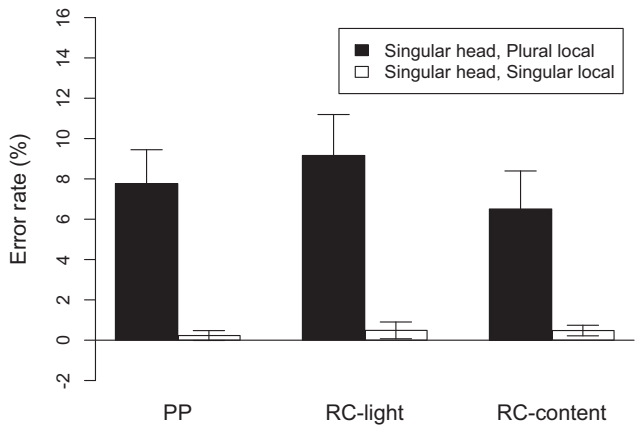

Figure 2. Experiment 2 untransformed agreement error rates as a function of modifier and noun number. Error bars indicate ±1 standard error of the mean, computed by items. PP = prepositional phrase; RC = relative clause.

|*t*|s < 2.20, *p*s > .20). In the ANOVA analyses, the noun number main effect reached significance by participants but not by items.

Discussion

Like in Experiment 1 and nearly all other studies in the literature, a large noun number effect was observed, with larger agreement error rates when the local noun was plural than when it was singular (e.g., Bock & Miller, 1991; Eberhard, 1997; Eberhard et al., 2005). However, modifier and noun number did not interact, with the size of the mismatch effect for PPs not statistically different from that for either RC condition. Thus, the current study replicated Experiment 1 (with more power) and provided no evidence for structural effects on agreement when other differences between PPs and RCs were minimized.

The second goal of this experiment was to determine whether the presence of content verbs in RC conditions could have been responsible for the reduced mismatch effects in Bock and Cutting's (1992) and Solomon and Pearlmutter's (2004) RC conditions; but the lack of a reliable difference in mismatch effects between Experiment 2's PP and RC-content conditions (and between its two RC conditions) suggests that content verb presence was not the cause.[3]

General Discussion

Together, the results of Experiments 1 and 2 provide further support for an account of agreement production that does not involve a hierarchical component (also see Gillespie & Pearlmutter, 2011b). In Experiment 1, which varied modifier structure while controlling for other factors known to affect agreement computation, a large mismatch effect was observed, but the PP and RC mismatch effects were equal. These results suggested that agreement computation was not constrained by structure. Experiment 2 provided a direct replication of Experiment 1 and introduced an additional RC condition that contained content verbs to determine whether previously observed structural effects were due to differences in properties of the linking words used across structural conditions. Mismatch effects were equal across all three conditions, replicating the findings of Experiment 1 and suggesting that the mere presence of content verbs in RC conditions was not responsible for reducing the RC mismatch effect relative to the PP mismatch effect in Bock and Cutting (1992) and Solomon and Pearlmutter (2004).

We further discuss the implications of these results below, but there are at least two general concerns with the current evidence that must be considered. First, the tasks used in these studies only approximate the natural production process because they involve a comprehension component. To the extent that this is an issue for the current study, it is an issue for essentially all other studies in the literature because nearly all involve a version of the recall task (e.g., Bock & Cutting, 1992; Bock & Miller, 1991) or the no-recall task (e.g., Gillespie & Pearlmutter, 2011b; Solomon & Pearlmutter, 2004). Experiment 1 directly tested whether task had an effect on the size of mismatch effects, and no significant interactions were found, suggesting that the processing in both tasks is similar (see Gillespie & Pearlmutter, 2011b, for further discussion about how comprehension may influence production in these tasks). However, given that the scope of the planning hypothesis crucially relies on the relative timing of the planning of the elements to explain mismatch effects, it will be necessary to design paradigms that better approximate the natural planning process while reliably eliciting preambles with desired properties without requiring a comprehension component (see Gillespie & Pearlmutter, 2011a, and Haskell & MacDonald, 2005, for two possibilities).

The second potential concern is that in these studies the conclusions depend on certain null effects: the absence of various Noun Number × Modifier interactions. The failure to find these interactions cannot be the result of a lack of power: First, in both experiments, the noun number effect was clear, replicating earlier results. Second, the clause-boundedness pattern was robust in previous studies, and with the current studies' greater number of participants per list (26+, compared to 10 for both Bock & Cutting, 1992, and Solomon & Pearlmutter, 2004) and equal or greater number of items per condition (six for Experiment 1 and for Solomon & Pearlmutter, 2004; four for Experiment 2 and Bock & Cutting, 1992), it should have been replicable. Third, Experiment 2 provided a direct replication of Experiment 1's findings with a separate, larger group of participants,

[3] On the other hand, while no significant interaction was present, the RC-content mismatch effect was numerically smaller than the mismatch effects for the other two modifier types (see Figure 2), suggesting that content verb presence in the earlier studies' RC conditions could have been at least a contributing factor. Furthermore, Experiment 2's RC-content plural local noun condition was rated as more integrated than the corresponding PP condition, which could have slightly inflated the RC-content condition's mismatch error rate (the 0.17 difference on the 7-point integration scale was significant, though this corresponds to just a 1.5% difference in mismatch effect sizes, based on Solomon & Pearlmutter's, 2004, second meta-analysis). Also worth noting, however, is that the verbs in the RC-content RCs added 2.4 syllables relative to the PPs, and that might have influenced agreement error rates as well: On the scope of planning account, more intervening material between head and local nouns makes interference less likely to the extent that it lengthens the delay between the planning of the two elements (Gillespie & Pearlmutter, 2011b), but it is unclear whether differences in syllables in particular necessarily alter grammatical planning time. If additional syllables do matter (in the direction predicted by scope of planning), this would have artificially deflated the RC-content condition's mismatch effect relative to the PP condition's.

Table 6
Experiment 2 Agreement Error Rate Results (Weighted Empirical Logit Linear Regression)

Effect	By participants			By items		
	β	*SE*	t_1	β	*SE*	t_2
Noun number (SP)	.43	.10	4.33*	.73	.10	7.44*
Modifier (RC-light)	.06	.07	0.81	.18	.14	1.36
Modifier (RC-content)	.00	.07	−0.01	.07	.14	0.53
Noun Number × Modifier (RC-light)	.05	.14	0.32	.02	.14	0.16
Noun Number × Modifier (RC-content)	−.12	.14	−0.84	−.12	.14	−0.84

Note. For the noun number variable, the SP level was sum-coded +0.5 and SS was −0.5. The modifier variable was treatment-coded, with the level in parentheses as +1.0 and PP as the base level. Thus, for both variables, betas (βs) estimate the difference between levels in log-odds space. SP = singular head, plural local noun; RC = relative clause.
* $p < .001$.

using the task that had shown a numeric trend toward a clause-boundedness effect in Experiment 1.

Another possible approach for a structure-based theory to avoid having to account for the Experiments 1 and 2 results is to argue that the structural manipulation itself was ineffective or irrelevant. We consider a range of these possible arguments below, but one reason to doubt them is the combination of (a) the PP versus RC effects in Bock and Cutting (1992; and the Solomon & Pearlmutter, 2004, replication) were substantial, (b) earlier work (e.g., Eberhard et al., 2005; Franck, Lassi, Frauenfelder, & Rizzi, 2006; Franck et al., 2002) has treated those results as evidence of a structural effect on agreement processing, and (c) we know of no syntactic theory that would distinguish the RC structures in the earlier work from those in Experiments 1 and 2 or the PP structures in Bock and Cutting (1992) from those in the current experiments. While on some syntactic theories Solomon and Pearlmutter's (2004; Experiment 5) PPs might be argued to attach differently from those in the current experiments (see Solomon & Pearlmutter's discussion of argument vs. adjunct attachment), the difference would be in the direction of increasing the difference in hierarchical distance between the PPs and RCs in the current experiments, and the current experiments of course showed no effect of hierarchical distance at all. Thus, whatever the source of the difference between, on the one hand, (what appeared to be) clause-boundedness or hierarchical distance effects in Bock and Cutting and Solomon and Pearlmutter and, on the other, the absence of such effects in Experiments 1 and 2, that source cannot be structural. If the earlier results were considered evidence of clause-boundedness or hierarchical distance effects, the lack of appearance of the same patterns in Experiments 1 and 2 must be considered evidence against such effects.

Alternatively, a structure-based account might reject both the earlier results and the current experiments' as having insufficient or irrelevant structural manipulations (and thus relying on some other difference between the earlier and the current stimuli to explain the difference in effects). One suggestion would be that the PP versus RC contrast is only a matter of clause-boundedness, not hierarchical distance, and that structural constraints on agreement depend on the latter and not the former. However, as far as we can tell, every modern syntactic theory (e.g., Chomsky, 1981, 1995; Dalrymple, 2001; Pollard & Sag, 1994; Sag, 1997) requires additional structure in the RC cases: This is the structure specifically associated with the clausal material itself (e.g., the relativizer *that*), linking the top of the RC's VP to the RC's attachment point into the head NP's structure, and formalized as (for example) S and/or S′ nodes in early transformational grammar (Chomsky, 1965), RP in Pollard and Sag (1994), and the CP-IP complex in Chomsky (1981) and Dalrymple (2001). So the simple count of nodes through which an errant feature would have to pass in order to (incorrectly) exit the modifier—hierarchical distance—is necessarily larger in the RC conditions than in the PP conditions, although the exact value will vary with the choice of syntactic theory.[4] This difference is also the basis for Franck et al.'s (2002) argument that a hierarchical feature-passing account of agreement can explain Bock and Cutting's (1992) results.

A second alternative would be to accept the difference in hierarchical distance between the PP and RC cases but suggest that only certain syntactic nodes count for hierarchical distance. Feature-passing theories have not been explicit about this (see Solomon & Pearlmutter, 2004, for discussion), but two possibilities are that only maximal projections are relevant or that only nodes out of which agreement should not grammatically be allowed to spread are relevant. The maximal projection possibility fails because at least one of the extra nodes identified above for the RC's clausal material is always a maximal projection (and the PP node will be matched by the RC's VP node). For the second possibility, the NP node for the local noun might count as a node out of which agreement should not be able to spread grammatically, but it will count identically for PPs and RCs. The PP node in the PP condition might also count, but the VP node in the RC condition should count in the same way. For the RC, however, at least one of the clausal-material nodes must also be relevant because agreement is not grammatically permitted to spread outside of a clause. Thus the extra hierarchical depth in the RC conditions yields at least one extra layer of relevant embedding for feature-passing, just as in Franck et al. (2002).

[4] Arguments based on differences in the position of the modifier's attachment relative to the subject NP node fail as well: Most theories will postulate (identical) adjunct attachment of the Experiments 1 and 2 PPs and RCs; see also Solomon and Pearlmutter (2004) and Gillespie and Pearlmutter (2011b) for more detailed discussion of potential effects of modifier attachment height.

A final alternative would be to rely only on clause-boundedness and not hierarchical distance as a relevant structural factor for agreement and then to suggest that the difference in clause-boundedness in the current stimuli is insufficient to drive a difference in agreement error rates. While we cannot entirely rule this out, it appears incompatible with current structural accounts of agreement phenomena, which require a finer grained set of structural distinctions (e.g., Eberhard et al., 2005; Franck et al., 2004, 2006, 2002; Hartsuiker et al., 2001; Vigliocco & Hartsuiker, 2002). In eliminating the effect of a single clause boundary, it seems less than easily compatible with a host of results in the psycholinguistic literature demonstrating notable consequences of a clause boundary (relative to a phrase boundary) for (for example) boundary identification (e.g., Fodor, Bever, & Garrett, 1974), sentence memory (e.g., Jarvella, 1971), prosody (e.g., Schafer, Speer, Warren, & White, 2000), and processing of ambiguity (e.g., Carlson, Clifton, & Frazier, 2001; Garnsey, Pearlmutter, Myers, & Lotocky, 1997).

Assuming, then, that the Experiments 1 and 2 null effects are informative, and given that the syntactic manipulations involve both a clause boundary and nontrivial differences in hierarchical distance, structure-based theories cannot account for the current results, at least if they rely on clause-boundedness or hierarchical feature-passing as predictive factors. One alternative structural approach is the minimalist-syntax-based theory of Franck and colleagues (Franck et al., 2006; Franck et al., 2010; Franck, Vigliocco, Antón-Méndez, Collina, & Frauenfelder, 2008). This approach does not rely on either clause-boundedness or hierarchical feature-passing as factors; instead, it describes agreement computation as (in part) a product of a variety of processes derived directly from current syntactic theory. It critically makes use of the notion of intervention by an interfering element between a head noun and the syntactic target of the head's agreement features (typically a node that will link those features to the verb) at a particular point during a syntactic derivation. For stimuli like those in Experiments 1 and 2, because all involve intervening material that is internal to the subject NP, the factor that determines the relative degree of interference is the hierarchical depth of the local noun relative to the subject NP node (Franck et al., 2006, pp. 208–209),[5] and the prediction is identical to that for hierarchical feature-passing: More deeply embedded local nouns will interfere less, so mismatch effects should be smaller for RCs than for PPs. In fact, the bases for this part of the theory are the studies taken as evidence for hierarchical feature-passing (e.g., Bock & Cutting, 1992; Franck et al., 2002), so a different prediction for the Experiments 1 and 2 results would require some change or addition to the theory to account for the earlier results.

Unlike structure-based accounts, Gillespie and Pearlmutter's (2011b) scope of planning account does predict the lack of an interaction in Experiments 1 and 2: When semantic integration and linear distance between the head and local noun are equated (along with other factors that might affect their relative time of planning), interference in the form of mismatch effects should be equal. This account is also compatible with conceptual number and lexical and morphophonological effects (e.g., as in Eberhard et al., 2005), but these were also controlled in Experiments 1 and 2.

Additionally, the scope of planning account may be able to explain previous clause-boundedness effects, if one or more of the factors controlled in Experiments 1 and 2 but varying in the prior clause-boundedness experiments affects the relative timing of planning of the head and local noun. As noted above, Experiment 2 examined one such property, the presence of a content (vs. function) linking word, and while there was a slightly suggestive numeric pattern, this hypothesis did not appear to be particularly viable.

Gillespie and Pearlmutter (2011c) investigated two other candidate properties that differed between PP and RC conditions in previous studies: verb frequency and verb transitivity bias. In general, low-frequency words are processed more slowly than high-frequency words (Oldfield & Wingfield, 1965), and the linking verbs used in previous studies' RC conditions were much lower in frequency than the prepositions used in their PP conditions. This could have led to later planning of the local nouns in the RCs, placing them more often outside the scope of planning of the head noun. Similarly, the linking verbs in the previous studies' RC conditions varied in transitivity bias—how often they occurred with a direct object NP—while the prepositions in the PP conditions nearly always required a following NP. In sentence comprehension, more predictable words in a given context are processed faster than less predictable words (e.g., Garnsey et al., 1997; Levy, 2008; Trueswell, Tanenhaus, & Kello, 1993), and speakers are sensitive to verb bias during production as well, when similar verb bias measures have been shown to influence production choices (e.g., Jaeger, 2010; Stallings, MacDonald, & O'Seaghdha, 1998).

With the idea that higher frequency verbs and verbs with higher probabilities of being transitive might speed up processing of their objects, thus increasing the chance that the object noun and the head noun would be simultaneously active during planning, Gillespie and Pearlmutter (2011c) examined whether varying the frequency or transitivity of the verb in an RC modifier would influence mismatch effects. The experiments used RC-content stimuli like those in Experiment 2 but varied the verb in the RC based on frequency (e.g., *The farmer who pushed/poked the stubborn goat[s],* with *pushed* higher frequency than *poked*) and, separately, based on verb transitivity (e.g., *The actor who quoted/yelled the line[s],* with *quoted* strongly transitive). As in Experiment 2, however, we found no reliable differences in mismatch effects as a function of either of these properties. These findings suggest that differences in linking word properties between PP and RC conditions in previous studies were unlikely to have been responsible for the observed structural effects.

Another difference between the current and previous studies was that the general meaning of the PP and RC conditions was explicitly matched within items in the current experiments, whereas the general meaning of the PP and RC versions of an item tended to vary much more in previous studies. Depending on the nature of the meaning differences, this might be a substantial factor contributing to the PP versus RC difference.

Meaning has been hypothesized to affect agreement computation by affecting the conceptual number of individual words as

[5] Franck et al. (2004) did at one point (p. 155) suggest that encoding of the head and local noun will be "simultaneous" in cases involving subject-internal modifiers; but the authors eventually (p. 169) seemed to endorse the same theoretical approach as later work (involving hierarchical depth; Franck et al., 2006, 2010, 2008). How the "simultaneous encoding" proposal would predict any differences on its own for stimuli like those in Franck et al. (2002) is unclear.

well as the subject NP as a whole (see, e.g., Bock & Middleton, 2011; Eberhard et al., 2005; and references therein), and the variety of semantically rich content verbs used in previous RC conditions may have produced differences in conceptual number which reduced RC mismatch effects. For example, RCs with semantically rich content verbs may be more likely to be interpreted as restrictive than PPs or RCs with less semantically specific verbs (like those in Experiments 1 and 2). The potential increase in restrictiveness of RC modifiers with semantically rich content verbs might in turn bias the conceptual number of the referent toward the singular because the NP may be more likely to be interpreted as referring to a specific individual from a set of potential alternatives, with the restrictive modifier serving a similar function to a singularly marked quantifier (e.g., *One key to the cabinets* vs. *The key to the cabinets;* Eberhard, 1997). This possibility is an interesting avenue for further research on how meaning relations may affect agreement processing.

In sum, these studies suggest that earlier clause-boundedness effects may have been confounded with differences in meaning and with other properties that may affect the timing of planning. When such properties are controlled, clause-boundedness does not influence agreement error rates, and combined with Gillespie and Pearlmutter's (2011b; Experiment 1) results, which showed that degree of syntactic embedding of a local noun also does not influence agreement error rates, these results suggest that structural properties neither directly constrain agreement computation nor form the underlying mechanism for such computation. Instead, we argue that agreement computation is governed by lexical and conceptual factors, as well as by processing constraints related to memory and to the timing of planning. Future work will be necessary to determine the extent to which structural and semantic properties play independent roles in planning processes in language production.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, England: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language, 56,* 65–85.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59,* 457–474.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60,* 92–111.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31,* 99–127.

Bock, K., & Middleton, E. L. (2011). Reaching agreement. *Natural Language and Linguistic Theory, 29,* 1033–1069.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology, 23,* 45–93.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive yet high-quality data? *Perspectives on Psychological Science, 6,* 3–5.

Carlson, K., Clifton, C., Jr., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language, 45,* 58–81.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht, the Netherlands: Foris.

Chomsky, N. (1995). *The minimalist program.* Cambridge, MA: MIT Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dalrymple, M. (2001). *Syntax and semantics: Vol. 34. Lexical functional grammar.* San Diego, CA: Academic Press.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5,* 313–349.

DiBattista, A., & Pearlmutter, N. J. (2011). Effects of semantic integration on phrase and word ordering errors in production. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2295–2300). Austin, TX: Cognitive Science Society.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language, 36,* 147–164.

Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language, 41,* 560–578.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review, 112,* 531–559.

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language.* New York, NY: McGraw-Hill.

Franck, J., Cronel-Ohayon, S., Chillier, L., Frauenfelder, U. H., Hamann, C., Rizzi, L., & Zesiger, P. (2004). Normal and pathological development of subject–verb agreement in speech production: A study on French children. *Journal of Neurolinguistics, 17,* 147–180.

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition, 101,* 173–216.

Franck, J., Soare, G., Frauenfelder, U. H., & Rizzi, L. (2010). Object-interference in subject–verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language, 62,* 166–182.

Franck, J., Vigliocco, G., Antón-Méndez, I., Collina, S., & Frauenfelder, U. (2008). The interplay of syntax and form in sentence production: A cross-linguistic study of form effects on agreement. *Language and Cognitive Processes, 23,* 329–374.

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject–verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes, 17,* 371–404.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37,* 58–93.

Gibson, E., Piantadosi, S. T., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistic Compass, 5,* 509–524.

Gillespie, M., & Pearlmutter, N. J. (2011a). Effects of semantic integration and advance planning on grammatical encoding in sentence production. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.

Gillespie, M., & Pearlmutter, N. J. (2011b). Hierarchy and scope of planning in subject–verb agreement production. *Cognition, 118,* 377–397.

Gillespie, M., & Pearlmutter, N. J. (2011c, November). *Imageability in subject–verb agreement production*. Poster presented at the Fifty-second Annual Meeting of the Psychonomic Society, Seattle, WA.

Hartsuiker, R. J., Antón-Méndez, I., & van Zee, M. (2001). Object attraction in subject–verb agreement construction. *Journal of Memory and Language, 45,* 546–572.

Haskell, T. R., & MacDonald, M. C. (2005). Constituent structure and linear order in language production: Evidence from subject–verb agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 891–904.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61,* 23–62.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior, 10,* 409–416.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 824–843.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106,* 1126–1177.

Nicol, J. L. (1995). Effects of clausal structure on subject–verb agreement errors. *Journal of Psycholinguistic Research, 24,* 507–516.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology, 17,* 273–281.

Pearlmutter, N. J., & Solomon, E. S. (2007, March). *Semantic integration and competition versus incrementality in planning complex noun phrases.* Paper presented at the 20th Annual CUNY Conference on Human Sentence Processing, San Diego, CA.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar.* Chicago, IL: University of Chicago Press.

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org

Sag, I. A. (1997). English relative clause constructions. *Journal of Linguistics, 33,* 431–484.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research, 29,* 169–182.

Schneider, W. (1988). Micro Experimental Laboratory: An integrated system for IBM PC compatibles. *Behavior Research Methods, Instruments, & Computers, 20,* 206–217.

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology, 49,* 1–46.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language, 39,* 392–417.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 528–553.

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin, 128,* 442–472.

Appendix A

Experiment 1 Stimuli

The singular versions of the stimuli are shown below. The linking words used in the PP and RC versions are shown, separated by slashes. The plural, local noun versions were created by making the last noun plural.

1. The pizza with/that had the missing slice
2. The phone with/that had the new keypad
3. The truck with/that had the special bumper
4. The ship with/that had the spacious deck
5. The desk with/that had the sliding drawer
6. The shark with/that had the strong fin
7. The shirt with/that had the expensive fabric
8. The plant with/that had the delicious root
9. The stereo with/that had the tiny switch
10. The loaf with/that had the exotic grain
11. The telescope with/that had the polished lens
12. The television with/that had the sharp image
13. The fan with/that had the wide blade
14. The box with/that had the dented corner
15. The statue with/that had the imported stone
16. The beach with/that had the sloping dune
17. The hotel with/that had the luxury suite
18. The rollerblade with/that had the metal axle
19. The concert with/that had the rock band
20. The zoo with/that had the controlled habitat
21. The movie with/that had the famous scene
22. The episode with/that had the surprise ending
23. The satellite with/that had the integrated computer
24. The newsletter with/that had the insightful article

(*Appendices continue*)

Appendix B

Experiment 2 Stimuli

The singular versions of the stimuli are shown below. The linking words used in the PP, RC-light, and RC-content versions are shown (respectively) separated by slashes. The plural, local noun versions were created by making the last noun plural.

1. The pizza with/that had/that included the yummy topping
2. The phone with/that had/that included the new keypad
3. The truck with/that had/that included the special bumper
4. The cupcake with/that had/that held the birthday candle
5. The desk with/that had/that contained the sliding drawer
6. The bus with/that had/that contained the fancy bathroom
7. The shirt with/that had/that included the expensive button
8. The plant with/that had/that contained the rare enzyme
9. The stereo with/that had/that included the tiny switch
10. The loaf with/that had/that contained the exotic grain
11. The telescope with/that had/that included the polished lens
12. The television with/that had/that displayed the sharp image
13. The fan with/that had/that included the wide blade
14. The box with/that had/that contained the spare part
15. The sculpture with/that had/that contained the pretty fountain
16. The beach with/that had/that featured the sloping dune
17. The hotel with/that had/that featured the luxury suite
18. The rollerblade with/that had/that included the metal axle
19. The concert with/that had/that featured the rock band
20. The zoo with/that had/that featured the controlled habitat
21. The movie with/that had/that contained the famous scene
22. The episode with/that had/that included the surprise ending
23. The satellite with/that had/that included the modern computer
24. The newsletter with/that had/that contained the insightful article

Received December 24, 2010
Revision received May 12, 2012
Accepted May 16, 2012 ■