Prosody reflects semantic integration in sentence production

Maureen Gillespie[a], Neal J. Pearlmutter[b], Stefanie Shattuck-Hufnagel[c]

[a]University of New Hampshire, [b]Northeastern University,

[c] Massachusetts Institute of Technology

Mailing address:
10 Library Way, Conant Hall
Department of Psychology
University of New Hampshire
Durham, NH 03824

E-mail: mtc2@cisunix.unh.edu (Gillespie), pearlmutter@neu.edu (Pearlmutter), sshuf@mit.edu (Shattuck-Hufnagel)

Running head: PROSODY REFLECTS SEMANTIC INTEGRATION

July 27, 2015 DRAFT, comments welcome. Please do not cite or quote without permission.

Abstract

Semantic integration, the degree of conceptual linkage between elements within an utterance, has been hypothesized to influence the timing of planning of elements within a phrase, such that highly semantically integrated elements are planned with more overlap than less integrated elements (Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004). But the evidence that integration affects the temporal separation of less integrated elements has been indirect. Word durations and the presence of prosodic breaks were examined in spoken responses from two agreement error elicitation studies that manipulated semantic integration (Gillespie & Pearlmutter, 2011, Exp. 2; Solomon & Pearlmutter, 2004, Exp. 4) to determine whether decreased semantic integration was associated with increased prosodic separation. All analyses provided evidence that speakers were more likely to prosodically separate elements that were weakly semantically integrated compared to items that were more strongly semantically integrated. These findings strengthen the hypothesis that prosodic timing may, at least in part, reflect underlying planning processes in sentence production.

Keywords: sentence production; linguistic planning; semantic integration; prosody; ToBI; word duration

The study of language production is concerned with how speakers translate non-verbal thoughts into meaningful, grammatical utterances. Models of language production generally separate the planning process into different levels. ? (?) separate the planning process into three main levels: the message level, which represents the speaker's intended meaning; the grammatical encoding level, which translates the meaning into a sequence of words; and the phonological encoding level, which translates the sequence of words into the sounds required to produce the utterance. The processing related to this translation process is often referred to as *planning*, and planning in production has been hypothesized to be cascading. In other words, as elements are being planned at one stage, planning at other levels is proceeding simultaneously. In addition, it is possible that multiple elements are simultaneously active at any given stage of production (?, ?).

Prosody generally refers to the rhythm, pitch, and timing of speech. One particularly interesting aspect of prosody is that it is affected by information at all stages of the production process. The intended meaning of a string of words can be conveyed through its intonational phrasing, suggesting that message level properties are reflected in prosodic structure. For example, if *John went to the store* is produced with a drop in pitch toward the end, this sentence will be interpreted as a statement; however, if the pitch rises at the end of the sentence, it is likely to be interpreted as a question (?, ?). Intonational breaks and lengthening are likely to appear at major syntactic boundaries (?, ?), suggesting prosody is sensitive to aspects of grammatical encoding, though syntactic units are often confounded with message units and prosodic phrasing does not always reflect syntactic constituent structure (?, ?). Additionally, prosody is affected by phonological encoding, as the intrinsic length of phonemes within a word affects the relationship between the duration of that word and its surrounding pauses (?, ?).

In addition to the influences described above, recent evidence suggests that some aspects of prosody reflect effects of real-time linguistic planning (for a review see ?, ?). Linguistic elements that are predictable within their context are shorter than less predictable elements (e.g., ?, ?, ?, ?, ?, ?, ?, ?). Speakers tend to lengthen words or produce disfluencies before material that is difficult to access, suggesting that speakers can use this increased time to aid in planning that material (?, ?, ?, ?). Intonational phrases, which are perceptually salient units of prosodic grouping, may reflect planning processes in production. For example, intonational phrases have been hypothesized to reflect planning units in production, with speakers appearing to be less likely to prosodically separate constituents that are semantically linked with intermediate phrase boundaries and intonational breaks (?, ?, ?).

Following up on work examining where intermediate phrase boundaries and full intonational boundaries[1] are likely to be produced, ? (?) developed a model of intonational boundary placement that assumed the purpose of intonational boundaries is to provide speakers with time to (1) recover from expending resources after planning and producing long constituents, and (2) plan upcoming material (see also ?, ?, ?). ? argue that planning units in production may be based on semantic relationships, as elements that are semantically and syntactically dependent upon each other (e.g., a verb and its argument(s)) often appear in the same intonational phrase (e.g., ?, ?, ?). ? (?) found that speakers were less likely to place intonational boundaries between the head of a verb phrase and its syntactically obligatory arguments (e.g., *investigated the crash*) than its nonobligatory adjuncts (e.g., *arrived after the crash*). They suggested that heads and their following constituents are more likely to be planned simultaneously when the following constituent is obligatory because heads and their obligatory arguments share a tight semantic link (see also ?, ?). Watson and colleagues' models of intonational boundary placement, which suggest a moderately incremental planning system with planning units partially specified by the semantic content of the utterance, predict intonational boundary placement as well as, and in some cases better than, older models that relied on syntactic structure (?, ?, ?).

? (?) suggested that the semantic integration effects observed in ? (?) are further evidence that semantic properties affect planning in production. ? claimed that elements that are more tightly conceptually linked are planned with more temporal overlap. ? termed the degree of conceptual linkage *semantic integration* and hypothesized that highly semantically integrated elements within

[1] For the purposes of this discussion, we will group these two types of boundaries and refer to them as intonational boundaries because both are associated with changes in F0 through accenting or boundary tones.

an utterance are more likely to be planned simultaneously[2], thus allowing their features to interfere with each other resulting in speech errors (also see ?, ?, ?, ?).

? (?) manipulated the degree of semantic integration in sentence preambles consisting of a head noun and a prepositional phrase (PP) modifier, using a subject-verb agreement error elicitation task (?, ?). Four versions of each stimulus item were created, in (1), by varying the degree of semantic integration between the head and local noun (*chauffeur* and *actor(s)*, respectively, as in (1)) as well as the local noun's number marking. (The code following each preamble indicates noun number for the head noun (N1) and the local noun (N2), with S meaning singular and P plural.) In integrated conditions (1a-b), the head noun and local noun were linked with the preposition *for* and were in a functional relationship, and in the unintegrated conditions (1c-d) the head noun and local noun were linked with *with* and were in an accompaniment relationship[3]. In agreement error elicitation studies, the size of mismatch effects are examined as an indicator of interference from local nouns. The mismatch effect is calculated by subtracting the agreement error rate for cases with a singular local noun (e.g., 1a, 1c) from the agreement error rate of the corresponding case with a plural local noun (e.g., 1b, 1d).

(1) a. The chauffeur for the actor (SS)
b. The chauffeur for the actors (SP)
c. The chauffeur with the actor (SS)
d. The chauffeur with the actors (SP)

? found that mismatch effects were larger in integrated conditions than in unintegrated conditions. They suggested that the mechanism of this integration effect was the relative timing of planning of the nouns within the preamble and suggested that the degree to which phrases are

[2] The mechanism by which semantic integration affects timing of planning does not specify whether high integration causes elements to be planned with a greater degree of temporal overlap than a baseline, or whether low integration causes elements to be planned with more temporal separation than a baseline. Thus, the only predictions that can be made about semantic integration and timing of planning are relative, with higher integration resulting in relatively more temporal overlap and lower integration resulting in less temporal overlap.

[3] Semantic integration ratings of these items confirmed that the versions with a functional relationship between the head and local noun (i.e., *for* conditions) were more integrated than versions with an accompaniment relationship between the head and local noun (e.g., *with* conditions). For additional examples and an analysis of the potential syntactic differences among integration conditions, see ? (?).

planned or activated together affects subject-verb agreement error rates, with increased planning overlap leading to larger error rates. These findings suggest that conceptual-level factors influence the timing of planning of elements within phrases.

? (?) extended ?'s (?) hypothesis that the degree of overlap in planning determines the size of mismatch effects in subject-verb agreement error production. They hypothesized that a local noun's planning distance from the head noun due to both linear proximity in the output utterance and semantic integration determines the size of interference effects, and that these effects are independent of hierarchical syntactic structure.

? (?) tested whether linear distance to the head noun and semantic integration combine to influence agreement error production, using stimuli like that shown in (2), containing a head noun (*book*; N1) followed by two PP modifiers that both modified the head noun, each containing a local noun (*page(s)*, *pen(s)*; N2 and N3). Under a combined linear distance and semantic integration account, the likelihood of interference would be a function of whether the interfering element was within the scope of planning of the head noun; only local nouns planned close enough in time to the head would create mismatch effects, with both decreased head-local linear distance and increased head-local semantic integration increasing the chance of overlap in planning.

The number of N2 and N3 was varied, as in (2). One PP (e.g., *with the torn page(s)*) was tightly integrated with the head noun, while the other (e.g., *by the red pen(s)*) was weakly integrated. Linear distance was manipulated by having two local nouns (N2 and N3) at different distances from the head noun, and switching the order of the PPs allowed control of semantic integration at each linear position. The versions in which N2 was tightly integrated with the head noun, as in (2a-d), were referred to as the early-integrated conditions, and the versions in which N3 was tightly integrated with the head noun, as in (2e-h), were referred to as the late-integrated conditions.

(2) a. The book with the torn page by the red pen (SSS)

b. The book with the torn pages by the red pen (SPS)

c. The book with the torn page by the red pens (SSP)

d. The book with the torn pages by the red pens (SPP)

e. The book by the red pen with the torn page (SSS)

f. The book by the red pens with the torn page (SPS)

g. The book by the red pen with the torn pages (SSP)

h. The book by the red pens with the torn pages (SPP)

Error rates in ? (?) showed a pattern reflecting a combination of linear distance to the head and semantic integration. The N2 mismatch effect (2b vs. 2a) was larger than the N3 mismatch effect (2c vs. 2a) in the early-integrated conditions, while the N2 and N3 mismatch effects were equal in the late-integrated conditions and both smaller than the N2 mismatch effect for the early-integrated conditions. These results point toward an account of agreement production that relies on scope of planning such that plural local nouns that are planned close in time to the head (due to linear distance and semantic integration) are the most likely to cause interference during agreement computation.

Additional evidence that semantic integration affects the timing of planning of elements within an utterance comes from exchange error patterns. In ? (2007; see also ?, ?) speakers described simple pictures using two noun phrases (NPs) linked by either a preposition (e.g., *in*, *on*, *above*) or the conjunction *and*. Half of the pictures depicted an integrated scene (e.g., a spot on an apple) and the other half depicted an unintegrated scene (e.g., an airplane flying over a cloud). In the critical trials, speakers were presented with a preposition to use in their descriptions, and speakers were more likely to exchange the NPs in the integrated conditions (e.g., *the spot on the apple* → *the apple on the spot*) than in the unintegrated conditions (e.g., *the airplane above the cloud* → *the cloud above the airplane*). ? followed Solomon and Pearlmutter (2004) in suggesting that integrated elements are planned with more temporal overlap, which increases the chance of the later-intended element being incorrectly placed in the earlier position.

A critical question about language production that is raised by these findings is how far in advance speakers plan parts of their utterance, and how this planning unfolds as an utterance is produced. Unfortunately, existing methods for examining planning in language production have some limitations. Traditionally, the way to study planning in production has been to examine distributions of errors elicited experimentally or through an observed corpus (e.g., ?, ?, ?, ?, ?, ?). But it is well known that there may be biases in observation when collecting a corpus (?, ?, ?), and eliciting errors in the lab can be difficult. Even when errors are obtained in a laboratory setting they may be so rare that the sparse distributions can be troublesome for many methods of data analysis. Another fairly standard way of examining how far in advance speakers plan parts of their utterances is using speech onset times (e.g., ?, ?, ?); however, speech onset times only reflect the planning that happens prior to initiating speech. Given that planning in language production is thought to be at least somewhat incremental, these speech onset time measures do not take into consideration planning that happens as an utterance is unfolding. Finally, eye-tracking has recently been used to study ongoing planning during production, but nearly all studies to date have shown evidence of radical incrementality (i.e., words are planned one-by-one, with little or no simultaneous planning) in the production system (e.g., ?, ?, ?, ?, ?, ?), which seems unlikely given that speakers often produce speech errors where they exchange words and sounds across fairly large distances during spontaneous production (?, ?, ?; see ?, ?, and ?, ?, for related discussion). In short, there is a fairly limited inventory of methods for studying timing during language production, and this inventory has produced incomplete and sometimes conflicting data. ? (?) suggest that analyzing prosodic timing and phrasing may be as informative as analyzing speech errors when investigating questions about ongoing planning during sentence production.

Currently, the only evidence that semantic integration affects timing in production comes from speech error data (?, ?, ?, ?), and errors are at best an indirect measure of timing. Thus, the goal of these studies was to examine whether semantic integration affects timing during production, using more direct measures of timing. If semantic integration affects timing of planning in language production, differences in timing of planning are reflected in differences in timing of articulation (e.g., ?, ?), there should be more temporal separation in the output utterance between less integrated elements than between tightly integrated elements. One way temporal separation of unintegrated elements could arise in the output utterance is through the production of acoustically/durationally marked breaks, with these being less likely to be produced between tightly semantically linked elements (see ?, ?, ?). An additional, or alternative, way temporal separation due to timing of planning differences associated with semantic integration may arise is through an increase in the overall time elapsed between less integrated (vs. more integrated) elements in the output utterance, with this increase in separation arising from increased word durations and/or longer pauses.

To examine these possibilities, recordings from two agreement error elicitation studies that manipulated semantic integration were obtained (?, ?, ?). Break indices using ToBI (?, ?) and word durations were examined to determine if less integrated elements were more likely to be temporally separated than more integrated elements.

Experiment 1

The stimuli used in Experiment 1 were from ?'s (2011) Experiment 2, as exemplified in (2). All critical stimuli were subject noun phrases (NPs) containing a head NP (*the book*) and two PP modifiers (PP1, PP2; *with the torn page(s)*, *by the red pen(s)*). One PP modifier was an attribute of the head NP (*with the torn page(s)*) and the NP PP pair was rated as tightly semantically integrated (see Table 1). The other PP modifier expressed a locative relationship to the head NP (*by the red pen(s)*, and was rated as less semantically integrated (see Table 1). The order of the PPs was manipulated, with PP1 always appearing adjacent to the head NP.

In ?'s (2011, Exp. 2), in the early-integrated condition agreement errors were likely to occur when N2 was plural but very unlikely to occur when N3 was plural. In the late-integrated condition, agreement errors were equally likely when N2 or N3 was plural. Assuming an incremental planning system that is affected by meaning relations (?, ?, ?, ?), the most straightforward

Table 1: Gillespie and Pearlmutter's (2011; Exp. 2) Stimuli, Attachment Preferences, and Semantic Integration Ratings by Condition

		Sentence Region / Word Position												
		NP		PP1				PP2					Semantic Integration Rating	
	Noun											%N1		
PP Order	Number	D1	N1	P1	D2	A1	N2	P2	D3	A2	N3	Attachment	N1-N2	N1-N3
Early-integrated	SPP	The	book	with	the	torn	pages	by	the	red	pens	98.0 (5.1)	5.75 (.60)	2.12 (.54)
	SPS	The	book	with	the	torn	pages	by	the	red	pen	98.9 (6.1)	5.70 (.56)	2.16 (.62)
	SSP	The	book	with	the	torn	page	by	the	red	pens	97.6 (5.5)	5.63 (.63)	2.21 (.60)
	SSS	The	book	with	the	torn	page	by	the	red	pen	96.4 (8.2)	5.74 (.74)	2.24 (.62)
Late-integrated	SPP	The	book	by	the	red	pens	with	the	torn	pages	99.2 (3.3)	2.19 (.56)	5.68 (.71)
	SPS	The	book	by	the	red	pens	with	the	torn	page	98.9 (2.7)	2.03 (.47)	5.71 (.64)
	SSP	The	book	by	the	red	pen	with	the	torn	pages	97.7 (4.0)	2.18 (.58)	5.57 (.75)
	SSS	The	book	by	the	red	pen	with	the	torn	page	97.7 (3.8)	2.25 (.65)	5.63 (.66)

Note. Standard deviations are in parentheses. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked). For word positions, D = Determiner; N = Noun; P = Preposition; A = Adjective. The number column shows the number marking of each of the three nouns, in order, with S = singular and P = plural.

set of predictions concerning the production of acoustically/durationally marked breaks follows from the error patterns observed in ? (?): In the early-integrated conditions very few acoustically/durationally marked breaks should be produced between NP and PP1 because N2 should be planned overlappingly with N1, while many acoustically/durationally marked breaks should be observed after PP1, as this would separate N3 from the head noun (N1). In the late-integrated conditions, the number of acoustically/durationally marked breaks produced after the head NP and PP1 in the late-integrated conditions should be approximately equal.

Semantic integration might also be reflected in the prosodic timing of these utterances. If semantic integration affects timing of planning (?, ?, ?), more integrated elements may be less temporally separated in the output utterance than less integrated elements, predicting that in these preambles words produced between tightly integrated elements should be shorter than words produced between less integrated elements. Thus Experiment 1 examined whether the durations of words linking less integrated head and local nouns (i.e., N1 and N2 in the late-integrated conditions) were longer than the durations of words linking more integrated head and local nouns (i.e., N1 and N2 in the early-integrated conditions). Additionally, durations of words in PP2 were examined. Though it is unclear how durations may be affected when a full PP separates the head noun and the second local noun, the predictions outlined for PP1 can be applied to PP2. Thus, durations linking less integrated head noun and local nouns (i.e., N1 and N3 in the early-integrated conditions) should be be longer than durations of words linking more integrated head and local nouns (i.e., N1 and N3 in late-integrated conditions). Various other measures were included in the duration models to control for factors that may affect word duration in addition to semantic integration.

Method

Participants

Recordings from 24 speakers (3 from each presentation list) in ? (?) Experiment 2 were used in the current experiment. The speakers were chosen to be largely fluent (i.e., they did not produce many trials with overt disfluencies), and 14 were male. Thus, the data in the current experiment were obtained by creating additional measures from the original recordings. ToBI analyses included data from 16 speakers (2 from each presentation list), and word duration analyses included data from all 24 speakers.

Materials

? (2011; Exp. 2) created 40 stimulus sets like that shown in Table 1 for their agreement error elicitation task. Each began with a head NP containing the determiner *the* (D1) and a head noun (N1), followed by two prepositional phrase (PP) modifiers, each of which consisted of a preposition (P1, P2), the determiner *the* (D2, D3), an adjective or modifier noun (A1, A2), and a local noun (N2, N3). One PP always described an attribute of the head noun using the preposition *with* (e.g., *with the torn page*), while the other PP specified a location for the head noun and used a locative preposition (e.g., *by the red pen*). The eight different versions of an item were created by varying N2 and N3 number and PP order, as shown in Table 1. Participants in ?'s (?) Experiment 2 read preambles aloud off a computer screen (preambles were presented for 50 ms/character) and then went on to complete the preambles as full sentences in any way that they chose. The table also shows the individual word and phrase labels used for reporting the results of the current experiment.

The stimuli were constructed so that PP2 reliably attached high to and modified N1, and this was confirmed by norming (see Gillespie & Pearlmutter, 2011, for details). Table 1 shows the normed mean percentage attachment of PP2 to N1 by condition; attachment preference did not vary with PP order.

ToBI labeling

A total of 640 tokens from 16 of the 24 speakers were obtained. ***** some of the following is repeated in Results, and the proportions in Table 4 are just the collapsed-over-labeler versions of Table 2 ***** In total, 509 tokens that did not contain preamble repetition errors or disfluencies were labeled with break indices using a slightly modified version of the ToBI coding system (?, ?). Two trained ToBI coders who were blind to the hypothesis completed all labeling, with each token coded by a single labeler. One coder labeled 54.2% of the tokens and a second coder labeled the remaining 45.8% of the tokens. Table 2 presents the distribution of labels at the end of each sentence region (defined by major syntactic phrases) used by each coder. One participant's data was coded by both labelers, and inter-rater reliability of ToBI labels was high (ICC$(3,2) = 0.94$, $p < .001$; ?, ?).

As in the standard ToBI coding system, break indices of 4 indicated intonational phrase boundaries (end-of-sentence intonation), break indices of 3 indicated intermediate phrase boundaries (lengthening with a boundary tone), break indices of 1 indicated normal inter-word boundaries, and break indices of 0 indicated that there was no discernible break between words. Though break indices of 2 are widely assumed to arise when there is duration lengthening on a word that is not associated with an intonational constituent (?, ?), there are no standard conventions for how break indices of 2 are defined in ToBI (see ?, ?). Additionally, there is some disagreement about the prevalence of breaks receiving break indices of 2 in labeled corpora (?, ?, ?). Because of the lack of

Table 2: Experiment 1 Percentage of ToBI Break Indices by ToBI Labeler, PP Order, and Sentence Region

			ToBI Break Index			
Labeler	PP Order	Region	0	1	2	3
Labeler 1	Early-Integrated	NP	0.0	95.5	3.7	0.7
		PP1	0.0	43.3	48.5	8.2
		PP2	0.0	6.7	18.7	74.6
	Late-Integrated	NP	0.0	96.5	3.5	0.0
		PP1	0.0	50.0	45.1	4.9
		PP2	0.0	7.0	12.7	80.3
Labeler 2	Early-Integrated	NP	0.0	98.3	0.9	0.9
		PP1	0.0	67.0	27.0	6.1
		PP2	0.0	7.0	9.6	83.5
	Late-Integrated	NP	0.0	99.2	0.8	0.0
		PP1	0.8	83.9	9.3	5.9
		PP2	0.0	5.9	9.3	84.7

Note. Break indices are applied to the end of the indicated region. NP = head noun phrase; PP1 = first prepositional phrase; PP2 = second prepositional phrase.

convention for labeling break indices of 2, labelers may not pay as much attention to acoustic markers that do not correspond to intonational constituent boundaries. In the current stimuli, we were interested in the patterns of prosodic separation at sentence-internal positions. Given that prosodic breaks and boundaries defined by ToBI as 3s and 4s were unlikely to occur in these contexts, we expected that the breaks observed in the preambles would be more subtle. Thus, conventions for labeling break indices of 2 were established for the current analyses. Hopefully establishing these conventions will provide a step toward better understanding of the perceptual experience associated with break indices of 2 in ToBI. Additional issues surrounding the categorization of 2s in ToBI will be discussed in the General Discussion.

Four prosodic cues were identified that may signal prosodic breaks: (1) lengthening, (2) pause insertion, (3) change in f0, and (4) pitch reset. Duration lengthening had to be perceptually salient. Pause insertion had to be perceptually salient as well as observable in the spectrogram. Change in f0 (usually a fall in pitch, in these declarative utterances) had to be perceptually salient and observable in the pitch track. Pitch reset was identified when the f0 following a word boundary returned to an equal or higher f0 than the preceding accented word (H* or !H*); the pitch reset had to be perceptually salient and observable in the pitch track. For a boundary to be assigned a break index of 2, one of these prosodic cues needed to be present *or* two or more only perceptually salient prosodic cues needed to be present.

Word duration measure

A total of 960 tokens were obtained from the full set of 24 speakers. Word boundaries for each word position within the preambles were hand-marked in Praat (?, ?) by one trained labeler who was blind to the hypothesis. Using the guidelines established in ? (?), reliable acoustic landmarks of phonetic segments beginning and ending words were located. The time between the beginning and ending landmarks was used as the word duration measure. Analyses using mixed-effect models predicting duration from semantic integration and a variety of control predictors were conducted. There were three types of control predictors included in the models: (1) speech rate, (2) phonological context, and (3) accessibility/predictability. The exact controls included in each model are indicated by check marks in Table 3 and described in detail below.

Table 3: Control Predictors Included in each Individual Word Position Model in Experiment 1

		Word Position									
Control Type	Predictor	D1	N1	P1	D2	A1	N2	P2	D3	A2	N3
Speech Rate	Rate	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Phonological Context	Length(w)		✓	✓		✓	✓	✓		✓	✓
	Length($w-1$)			✓	✓		✓	✓	✓		✓
	$w+1$ Vowel Initial	✓	✓		✓	✓	✓		✓	✓	
	$w-1$ Vowel Final			✓			✓	✓			✓
	Pause following w		✓				✓				✓
Accessibility/Predictability	Length($w+1$)	✓	✓		✓	✓	✓		✓	✓	
	F($w+1$)	✓	✓		✓	✓	✓		✓	✓	
	F($w-1$)			✓	✓		✓	✓	✓		✓
	F(w)		✓	✓		✓	✓	✓		✓	✓
	P($w+1 \mid w-1, w$)		✓	✓	✓	✓	✓	✓	✓	✓	
	P($w+1 \mid w$)	✓									
	P($w \mid w-1, w+1$)		✓	✓	✓	✓	✓	✓	✓	✓	
	P($w \mid w-1$)										✓
	P($w \mid w+1$)	✓									

Note. A check mark (✓) indicates that the control predictor was included in a particular model. D = Determiner; N = Noun; P = Preposition; A = Adjective; w = current word; $w-1$ = previous word; $w+1$ = next word; Length = length in phonemes; Vowel Initial = begins with a vowel; Vowel Final = ends with a vowel; F = log frequency; P = log probability.

Semantic integration predictor

The main predictor of interest was semantic integration condition (treated a two-level categorical predictor: early-integrated vs. late-integrated). As described above, in the early-integrated conditions PP1 was more integrated with the head noun (N1) and PP2 was less integrated with N1. The opposite was true for the late-integrated conditions. Mean semantic integration ratings (on a 1−7 scale) by condition are shown in Table 1.

Speech rate control predictor

Faster speech is characterized by shorter word durations and shorter pauses (e.g., ?, ?, ?) and is associated with the production of reduced word forms (e.g., ?, ?, ?). Thus it was important to control for a speaker's average speech rate as it is likely to affect their word durations.

Speech rate (*Rate*) Recordings of five filler trials from ? (?, Experiment 2) were obtained for each speaker. These fillers had a variety of structures and these structures were not identical to the critical items. Word durations for each word within the filler preambles were obtained using the same technique as was used for critical trials. Each speaker's log speech rate in syllables per second was included as a control predictor (for other duration models including a speech rate control predictor, see ?, ?, ?, ?).

Phonological context control predictors

The phonological context in which a word appears can affect its duration. For example, determiners appearing before adjectives and nouns that begin with a vowel are often produced in their full form (e.g., *the* rhymes with *pea*), while determiners preceding words that begin with a consonant are usually produced in their reduced form (e.g., *the* rhymes with *duh*; ?, ?, ?, ?). Word durations can also be affected by where in a prosodic phrase a word appears. The duration of words is likely to be longer when those words appear at the edge of an intonational boundary as words preceding intonational breaks often undergo phrase-final lengthening (?, ?, ?). Thus, it is important to control for various aspects of phonological and prosodic context.

Length of current word (*Length (w))* The words in the preambles differed among items and versions. Thus, the models had a control for the length of the word in phonemes, as words with more phonemes are likely to have longer durations.

Length of previous word (*Length (*w−1*)*) It is unclear whether speakers may show recovery effects (see ?, ?) after planning and producing phonologically long and complex material, thus the length of the immediately previous word in phonemes was included in models where the the length of the preceding word varied word.

Following word begins with vowel (w + 1 *Vowel Initial*) As noted above, determiners are more likely to be produced in their full form prior to a word beginning with a vowel. In addition, acoustic landmarks of vowels can be fairly difficult to locate (?, ?). Thus, this two-level categorical predictor was included to control for the possible differences in pronunciation when a word precedes a vowel.

Previous word ends with vowel (w – 1 *Vowel Final*) As mentioned above, acoustic landmarks indicating vowel offsets can be difficult to locate. Thus, this predictor was included in the models to control for possible variation of the acoustic landmarks indicating the offset of the vowel.

Pause following word (*Pause following w*) Speakers often paused after Noun2 and prior to producing the sentence completion (i.e., after Noun3). In the Noun models, the presence of a pause following the Noun was included as a two-level categorical predictor to control for lengthening effects that may be present due to prosodic break placement.

Accessibility and predictability control predictors

The principle of immediate mention (?, ?) states that speakers are likely to produce accessible (i.e., active) words as early in an utterance as possible (for evidence from optional *that*-mentioning see ?; ?, ?; for evidence from structural choices see ?, ?, ?, ?, ?, ?, ?), while less accessible elements are produced as late as possible in the utterance. Speakers often lengthen words or produce disfluencies (e.g., *um*, *uh*, restarts) if an upcoming word is not accessible, presumably to buy themselves time to plan the upcoming elements (?, ?, ?). Taken together, these findings suggest that speakers are sensitive to the accessibility of upcoming material and produce less accessible elements as late as possible. Thus, it is important to control for factors that may modulate accessibility of upcoming words independent of semantic integration.

The predictability of a word given its surrounding context has also been shown to affect word duration (e.g., ?, ?, ?, ?, ?) and structural choices (e.g., ?, ?, ?, ?) in natural speech, with more predictable elements showing reduced syntactic forms and shorter durations than less predictable words. Thus, the models included measures of lexical predictability within various contexts.

While accessibility and predictability have been shown to have independent effects on production (e.g., ?, ?, ?, ?), it is not the goal of Experiment 1 to determine how lexical accessibility and predictability affect word duration. Instead, the following controls were included in the models to ensure that differences in word duration predicted by the degree of semantic integration were not due to overall differences in lexical accessibility and predictability in the stimuli.

Length of following word (*Length (*$\mathbf{w+1}$*)*) The length of upcoming material has been shown to affect word duration. Speakers lengthen words before longer, more complex, and more difficult to access material (e.g., ?, ?, ?). This predictor was included in the models to control for differences in accessibility due to word length.

Frequency of current (*F(w)*), previous (*F(*$\mathbf{w-1}$*)*), and following (*F(*$\mathbf{w+1}$*)*) words Words that are more frequent in the language are named more quickly than lower-frequency words (e.g., ?, ?), suggesting that frequency affects lexical accessibility (e.g., ?, ?, ?). In addition, lower frequency words tend to be longer than higher frequency words (?, ?, cf. ?, ?). The frequency of an upcoming word can affect the duration of an earlier word, presumably due to processing needs associated with planning upcoming material (e.g., ?, ?, ?), and there is a possibility that the duration of following words could be affected by the frequency of a previous word if word durations adjust due to recovery processes (for a similar suggestion at the syntactic level see ?, ?). The log frequency of each word obtained from the SUBTLEXus database (?, ?) was included to control for duration differences associated with lexical accessibility in models where the word being modeled differed among items and versions.

Accessibility of following word (*P(*$\mathbf{w+1 \mid w-1,\ w}$*)*) Words are often lengthened if they precede less accessible material (?, ?). To control for accessibility of the word following the modeled word given its previous context, the log probability of the following word ($w+1$) given the two preceding words ($w-1$, w) was calculated from the Google n-gram corpus (?, ?). Since there were not two words of previous context for the Determiner1 model, accessibility of the following word was estimated using only Determiner1 as the preceding context. This predictor is referred to as $P(w+1 \mid w)$ in Table 3.

Predictability of word in its context ($P(\mathbf{w} \mid \mathbf{w}-\mathbf{1}, \mathbf{w}+\mathbf{1})$) A measure of predictability included in the models was the probability of a word occurring given the words it appeared between. The log probability of a word given the word that preceded and followed it was calculated from the Google n-gram corpus (?, ?). Because Determiner1 was not preceded by any previous context, the predictability measure for the Determiner1 model was estimated using the following word as its immediate context. This predictor is referred to as $P(w \mid w+1)$ in Table 3. Because Noun3 was not followed by any context, the predictability measure for the Noun3 model was estimated using the previous word as its immediate context. This predictor is referred to as $P(w \mid w-1)$ in Table 3.

Results

ToBI break index model

A total of 509 of the 640 tokens (79.5%) were included in the ToBI analysis after excluding tokens with disfluencies or preamble repetition errors. We focused on break indices after each noun within the preambles because breaks stronger than a 1 in ToBI were extremely rare in other positions. Counts of ToBI break indices by integration condition and sentence region are shown in Table 4. To determine if semantic integration had an effect on the presence of acoustically/durationally marked prosodic breaks at major syntactic boundaries within the preambles, a logistic mixed-effect regression model[4] was conducted predicting the presence of a break index greater than 1 (i.e., breaks stronger than those expected between words) as a function of semantic integration

[4] All analyses were conducted using R v. 2.15.0 (?, ?) and lme4 v. 0.999375-42.

Table 4: Experiment 1 Counts of ToBI Break Indices by PP Order and Sentence Region*** **clean up** ***

		ToBI Break Index			
PP Order	Region	0	1	2	3
Early-integrated	NP	0 (0.0)	241 (96.8)	6 (2.4)	2 (0.8)
	PP1	0 (0.0)	135 (54.2)	96 (38.6)	18 (7.2)
	PP2	0 (0.0)	17 (6.8)	36 (14.5)	196 (78.7)
Late-integrated	NP	0 (0.0)	254 (97.7)	6 (2.3)	0 (0.0)
	PP1	1 (0.4)	170 (65.4)	75 (28.2)	14 (5.4)
	PP2	0 (0.0)	17 (6.5)	29 (11.2)	214 (82.3)

Note. Percentages are in parentheses. NP = head noun phrase; PP1 = first prepositional phrase; PP2 = second prepositional phrase.

condition. Because break indices of 2 and 3 in ToBI are often associated with lengthening (see ?, ?), we decided to group these breaks together for the purposes of these analyses as the main question of interest is whether semantic integration affects timing in language production. The model included semantic integration, sentence region, and their interaction as fixed effects; and random intercepts for participants, items, and ToBI labelers. The semantic integration predictor was sum coded and the sentence region predictor was treatment coded with PP1 as the base level.

The untransformed percentage of break indices stronger than a 1 for each condition by sentence region is shown in Figure 1. The results of the regression model are shown in Table 5. Overall, participants produced evidence of more acoustically/durationally marked breaks after PP2 than after PP1, and speakers produced evidence of more acoustically/durationally marked breaks after PP1 than after NP. At PP1, more acoustically/durationally marked breaks were produced in the early-integrated conditions than in the late-integrated conditions. The integration effect at PP1 was

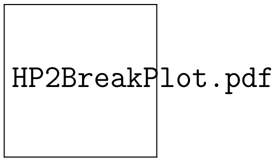

Figure 1: Percent breaks stronger than a 1 in ToBI as a function of position and semantic integration. Error bars are ±1 *SEM* computed by participants.

Table 5: Experiment 1 ToBI Results

Effect	β	*SE*	*z*-value
Semantic Integration (Early-int)	0.54	.20	2.72*
Position (NP)	−3.56	.31	−11.42**
Position (PP2)	3.42	.22	15.70**
Semantic Integration × Position (NP)	−0.22	.60	−0.37
Semantic Integration × Position (PP2)	−0.63	.42	−1.50

Note. The level shown in parentheses for the semantic integration variable was sum-coded +0.5 and the other level −0.5, so βs estimate the difference between the two levels of the variable at the base level of the Position variable (PP1). The level shown in parentheses for the position variable was treatment-coded +1 and compared to the base level (PP1), so βs estimate the difference between the level coded as +1 and the base level. Early-int = early-integrated; NP = head noun phrase; PP1 = first prepositional phrase; PP2 = second prepositional phrase.

**$p < .001$. *$p < .01$.

not significantly larger than at NP or at PP2; however, paired comparisons revealed that PP1 was the only sentence region where a significant integration difference was observed (NP: $z = 0.63$, $p = .53$; PP1: $z = 2.89$, $p < .01$; PP2: $z = -0.13$, $p = .90$).

Duration models

To determine the unique effect of semantic integration condition on each individual word position, separate linear mixed-effect regression models were conducted for each word within the preamble, controlling for speech rate, phonological context, and accessibility/predictability as shown in Table 3, and including random intercepts for participants and items. Of the 960 possible tokens, 680 contained no disfluencies or preamble repetition errors (70.8%) and were included in the following analyses. Tokens that were more than 3 SDs from the mean duration of the word being modeled were removed from the individual word models ($<$ 1% of the fluent tokens).

All predictors were standardized and fixed effect correlations were minimized by centering and residualizing. Any predictor in any individual model that was co-linear with the semantic integration predictor was regressed out of the final semantic integration predictor used in all models. After residualization, the fixed-effect correlations with the semantic integration predictor were low (*r*s $<$ .23). In the current study, we were not interested in how the control predictors affected word durations. We only included these predictors to ensure that any potential effect of the predictor of interest, semantic integration, were not due to other factors known to affect word duration; thus we do not report the coefficients for the control predictors and no measures were taken to ensure that other control factors were not co-linear with each other. The rule of thumb in regression modeling is to have at least 15 data points per predictor to avoid overfitting (?, ?). None of the following models violate this rule.

Standardized semantic integration coefficients for each duration model, with their 95% confidence intervals, are shown in Figure 2. A significant effect of integration condition was observed in the A1, D3, and A2 models. No other models showed a significant effect of integration condition ($|\beta|s < 0.141$; $|t|$s < 1.83, *p*s $>$ .06)[5].

[5] Additional analyses were conducted on only the tokens that did not result in a subject-verb agreement error being produced. Statistical patterns were identical to those reported below, except the Determiner2 model reached significance ($\beta = .12$; $t = 2.15$; $p < .05$).

Figure 2: The standardized semantic integration coefficient for each duration model in Experiment 1, with 95% confidence intervals obtained from MCMC sampling. A negative coefficient indicates shorter duration in the early-integrated conditions compared to the late-integrated conditions. The bars shown in dark gray indicate a significant effect of semantic integration condition on duration. SI = semantic integration.

Adjective 1 (A1)

After controlling for other factors, A1 durations were shorter in the early-integrated conditions than the late-integrated conditions ($\beta = -0.227$; $t = -6.06$, $p < .001$). The average A1 duration was 356ms, and with all else being equal, the model predicted a 284ms A1 in the early-integrated conditions and a 446ms A1 in the late-integrated conditions (a range 46% of A1's average duration).

Determiner 3 (D3)

After controlling for other factors, D3 durations were shorter in the early-integrated conditions than the late-integrated conditions ($\beta = -0.141$; $t = -2.45$, $p < .05$). The average D3 duration was 111ms, and with all else being equal, the model predicted a 96ms D3 in the early-integrated conditions and a 128ms D3 in the late-integrated conditions (a range 25% of D3's average duration).

Adjective 2 (A2)

A2 durations were longer in the early-integrated conditions than the late-integrated conditions ($\beta = 0.197$; $t = 5.09$, $p < .001$). The average A2 duration was 402ms, and with all else being equal, the model predicted a 489ms A2 in the early-integrated conditions and a 330ms A2 in the late-integrated conditions (a range 40% of D3's average duration).

Duration models including ToBI break indices

Additional duration analyses were conducted on the subset of fluent tokens that were labeled using ToBI to determine if the effects of semantic integration were conditional on whether a break was observed after PP1 (i.e., N2). All models contained the same control predictors as shown in Table 3, and a predictor indicating whether a break stronger than a 1 in ToBI was observed after PP1 was added to each model. The break presence predictor was allowed to interact with the semantic integration predictor. No models showed a significant interaction of integration condition and break presence, suggesting that the integration-related duration patterns observed did not differ depending on whether evidence for a break was produced. If a break stronger than a 1 in ToBI was identified following PP1 (i.e., N2), the duration of N2 was significantly longer ($t = 4.15$, $p < .001$). This pattern suggests that increased duration was a cue that labelers used in determining break strength, even when breaks were not associated with boundary tones (see ?, ?).

Discussion

The results of Experiment 1 suggest that semantic integration affects prosodic separation and timing. More breaks stronger than a 1 in ToBI were observed after PP1 in the early-integrated conditions than in the late-integrated conditions; thus, these breaks were more likely to be observed after a highly integrated PP1 than a less integrated PP1. If break patterns were to closely match the agreement error patterns observed in ?'s (2011) Experiment 2, breaks would have been least likely following early-integrated NPs and most likely following early-integrated PP1s, and breaks would have been equally likely to follow late-integrated NPs and PP1s, but overall less likely than after early-integrated PP1s. The break production patterns did not entirely match those predictions, but the effect of semantic integration on breaks produced after PP1 was in the predicted direction, with more breaks observed after early-integrated PP1s than late-integrated PP1s. The finding that breaks were rarely produced after NP is consistent with ? (?)'s planning and recovery hypothesis of prosodic break production, as speakers would be unlikely to need to recover after producing a single prosodic word.

In English, ambiguous structures that are intended to have high attachment are often preceded by prosodic boundaries (e.g., ?, ?, ?, ?, ?, ?, ?). In the preambles used in Experiment 1, PP2 modified the head noun, which resulted in high attachment. Thus it is possible that speakers often produced evidence for prosodic breaks immediately following PP1 to signal the high attachment of PP2. Experiment 1's semantic integration results suggest that semantic relations were responsible for break patterns, independent of attachment preferences, making speakers less likely to prosodically separate a tightly semantically-integrated PP2 from its head noun compared to a less-integrated PP2. However, the competing influences of syntactic structure (a preference for breaks to appear after PP1 to signal high attachment) and semantic relations (a dispreference to separate semantically-related material) could provide an additional explanation for the break patterns observed in the early-integrated and late-integrated conditions.

Watson et al. (?, ?, ?) found evidence that speakers were less likely to prosodically separate semantically related material, similar to Experiment 1; however, Watson et al. measured the presence of well-formed prosodic constituents (intermediate and full intonational phrases). The production of breaks creating well-formed prosodic constituents can be used to disambiguate globally ambiguous sentences (?, ?) and to indicate the appropriate grammatical structure of locally ambiguous utterances (e.g., ?, ?), suggesting that these breaks can be used to indicate particular grammatical and semantic relations a speaker has in mind (cf. ?, ?). In Experiment 1, the ToBI results were driven by the labeling of 2s between the two PPs. It is still unclear whether breaks that do not indicate well-defined prosodic constituents (e.g., the 2-labels observed in Experiment 1) serve the same purpose. We will return to this issue of processing-based and message-based explanations of prosodic variation in the General Discussion.

The word duration models showed significant effects of semantic integration in the predicted direction on A1 and A2 (the adjectives preceding the local nouns: N2 and N3, respectively); durations were shorter if the adjective was within the highly integrated PP compared to the less integrated PP, indicating that there was some duration adjustment associated with semantic integration. However, there was also an effect of semantic integration on D3 duration in the direction opposite the prediction, with shorter durations in the less integrated PP2 than in the early-integrated PP2. This effect may have at least partially canceled the effect of semantic integration on A2. Thus the predicted pattern of shorter durations associated with greater semantic integration was observed in PP1, but the results were not as clear in PP2 because the significant effects observed were in opposite directions. It is perhaps not surprising that results were less clear in PP2 because PP2 was separated from the head noun by a full prepositional phrase (PP1). Thus, there is the possibility that other factors may have influenced durations as the local nouns were more separated from the head noun.

? (?) hypothesized that increased semantic integration leads to increased planning overlap of the integrated elements; thus these findings support ?'s claim (also see ?, ?). ? (?) suggested that speakers produce breaks in order to recover after planning and producing semantically related material. Thus if the shorter duration between N1 and N2 that was observed in the early-integrated conditions indicates an increase in planning overlap, this finding provides an additional explanation for why more breaks were produced following PP1 in the early-integrated conditions than in the late-integrated conditions: Speakers needed to recover more after the more condensed planning of the head noun phrase and PP1 in the early-integrated conditions.

In sum, Experiment 1 showed clear evidence of prosodic break placement being affected by semantic relations within these NP PP PP preambles. Duration results showed patterns consistent with increased temporal separation of less semantically integrated elements; however, the duration patterns in PP2 were less clear and may have been affected by the preamble length or other semantic factors introduced by having an additional PP separating N3 from N1. Thus to more clearly investigate how semantic integration affects prosodic timing due to adjustments in word duration, additional duration analyses were conducted in Experiment 2 examining word durations in shorter preambles that did not include prosodic breaks, and that were also more closely matched at positions of interest.

Experiment 2

Experiment 1 showed that less semantically integrated elements were more likely to be temporally separated than tightly semantically integrated elements, but the duration analyses provided some contradictory results in the second prepositional phrase. To reexamine the duration findings observed in Experiment 1, similar duration analyses were performed on preambles from ?'s (2004; Experiment 4) stimuli that manipulated semantic integration and only contained a single prepositional phrase modifier following the head noun.

Semantic integration in the preambles in ? (2004; Experiment 4) was manipulated in a different way than in Gillespie and Pearlmutter's (2011; Experiment 2) preambles. Solomon and Pearlmutter used the same preposition (*with*) in all versions of their Experiment 4 preambles, and integration was manipulated by varying the local noun. The integrated versions reflected an attribute relationship between the head and local noun (e.g., *the sweater with the tiny hole(s)*) and the unintegrated versions reflected an accompaniment relationship between the head and local noun (e.g., *the sweater with the clean skirt(s)*). Thus stimuli were identical until the adjective preceding the local noun, making them better phonologically matched than those in Experiment 1.

The placement of acoustically/durationally marked breaks was preliminarily examined in a subset of these stimuli using the ToBI system (?, ?) modified as in Experiment 1. Break indices of 2 or 3 were often produced following Noun2, but in Experiment 2 the predictions concerned whether Noun1 and Noun2 were prosodically separated as a function of semantic integration. Just as in Experiment 1, there was little evidence of break indices stronger than a 1 within these preambles. We thus did not conduct further ToBI analyses; but the lack of such breaks indicates that these stimuli are well-suited to examining potential duration patterns.

This experiment therefore examined whether word durations linking unintegrated elements were longer than the duration of words linking integrated elements, using recordings of responses from ?'s (2004) Experiment 4. If semantic integration affects the temporal separation between words in these preambles, the durations of the words linking the head and local noun should decrease as integration increases.

Method

Participants

Recordings of all 24 critical items from 16 fluent speakers (4 from each presentation list, 7 males) were obtained, resulting in 384 total tokens.

Table 6: Solomon and Pearlmutter's (2004b; Exp. 4) Stimuli and Semantic Integration Ratings by Condition

		Word Position						Semantic
	Noun							Integration
Integration	Number	D1	N1	P1	D2	A1	N2	Rating
Integrated	SP	The	sweater	with	the	tiny	holes	5.53 (.63)
	SS	The	sweater	with	the	tiny	hole	5.44 (.80)
Unintegrated	SP	The	sweater	with	the	clean	skirts	3.28 (.79)
	SS	The	sweater	with	the	clean	skirt	3.25 (.89)

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun; D = Determiner; N = Noun; P = Preposition; A = Adjective.

Materials

Twenty-four stimulus sets like that shown in Table 6 were created by Solomon and Pearlmutter (2004; Exp. 4) for their agreement error elicitation task. Each began with a head NP (e.g., *The sweater*) followed by a PP modifier containing a local noun (e.g., *hole(s)*; *skirt(s)*) preceded by the determiner *the* and an adjective or noun modifier. The PP modifier always began with the preposition *with*. The integrated versions indicated an attribute relationship between the head and local noun, and the unintegrated versions indicated an accompaniment relationship between the head and local noun. The head noun was always singular, and the four different versions of an item were created by varying semantic integration and local noun number (see Table 6). Participants in ?'s (?) Experiment 4 read preambles aloud off a computer screen (preambles were presented for 40 ms/character) and then went on to complete the preambles as full sentences in any way that they chose. For the current analyses, each word position within the preamble was labeled as shown in Table 6.

Table 7: Control Predictors Included in each Individual Word Position Model in Experiment 2

		Word Position					
Control Type	Predictor	D1	N1	P1	D2	A1	N2
Speech Rate	Rate	✓	✓	✓	✓	✓	✓
Phonological Context	Length(w)		✓			✓	✓
	Length($w-1$)			✓			✓
	$w+1$ Vowel Initial	✓			✓	✓	
	$w-1$ Vowel Final			✓			✓
	Pause following w						✓
Accessibility/Predictability	Length($w+1$)	✓			✓	✓	
	F(N1)	✓	✓	✓	✓	✓	✓
	F(A1)	✓	✓	✓	✓	✓	✓
	F(N2)	✓	✓	✓	✓	✓	✓
	P($w+1 \mid w-1, w$)		✓	✓	✓	✓	
	P($w \mid w-1, w+1$)		✓	✓	✓	✓	
	P($w \mid w-1$)						✓
	P($w \mid w+1$)	✓					

Note. A check mark (✓) indicates that the control predictor was included in a particular model. D = Determiner; N = Noun; P = Preposition; A = Adjective; w = current word; $w-1$ = previous word; $w+1$ = next word; Length = length in phonemes; Vowel Initial = begins with a vowel; Vowel Final = ends with a vowel; F = log frequency; P = log probability.

Word duration measure

Word boundaries for each word within the preambles were hand-marked by the same research assistant as in Experiment 1, using the same procedures. This measure was used in models predicting duration from semantic integration and three types of control predictors, also as in Experiment 1: (1) speech rate, (2) phonological context, and (3) accessibility/predictability. The exact controls included in each model are indicated by check marks in Table 7 and described below.

Semantic integration predictor

The main predictor of interest was semantic integration. Semantic integration ratings, treated as a continuous predictor, for each version of each item, obtained from ? (?), were included in the model. Semantic integration was rated on a 1 to 7 scale with ratings closer to 7 indicating a tight conceptual link between the head and local noun and ratings closer to 1 indicating a weak conceptual link between the head and local noun. The mean semantic integration ratings for each version of the stimuli are shown in Table 6. In ? (?)'s original study, semantic integration was treated as a categorical predictor in the main agreement error analyses. Results were identical when semantic integration was treated as a categorical predictor (integrated vs. unintegrated) instead of a continuous predictor in the current analyses.

Speech rate control predictor

Speech rate (*Rate*) Recordings of six filler trials from ? (?, Experiment 4) were obtained for each speaker. These fillers had a variety of structures differing from the critical items. Speech rate was calculated as in Experiment 1.

Phonological context and accessibility/predictability control predictors

While the preambles examined in Experiment 2 were fairly well-matched in phonological properties, content words did vary across items and versions. In addition, speakers often placed pauses after the preamble. Thus, it is important to control for differences in duration that could be caused by phonological properties of the utterances. Accessibility/predictability control predictors were also included in the models, as in Experiment 1. For both types of control predictors, unless noted below, the predictors listed in Table 7 were handled identically to those in Experiment 1.

Pause following preamble (*Pause following w*) While only items with no overt pauses during the preambles (i.e., D1 – N2) were used in the analyses, speakers often paused after reading the preamble and prior to producing the sentence completion. In the N2 model, the presence of a pause following N2 was included as a two-level categorical predictor to control for lengthening effects that may be present due to intonational break placement.

Frequency of content words The log frequency of each content word obtained from the SUBTLEXus database (?, ?) was included to control for duration differences associated with lexical accessibility and length. These predictors are referred to as *F(N1)*, *F(A1)*, and *F(N2)* in Table 7.

Results

Twenty-nine of the 384 tokens were excluded because of preamble repetition errors and disfluencies, so that 355 tokens were included in the analyses (92.4% of all tokens). To determine if semantic integration affected word durations differently at different word positions, two linear mixed-effect regression models predicting log word durations were run and subjected to model comparison. A model that included integration, word position, and their interaction as fixed effect predictors, and participant and item intercepts as random factors, was compared to a model with identical parameters except only main effects of integration and word position. The model with the interaction term provided a significantly better fit ($\chi^2(5) = 19.94$, $p < .01$), indicating that the size of the integration effect differed across word position.

To determine the unique effect of semantic integration at each individual word position, separate analyses using linear mixed-effect regression models were conducted for each word, controlling for the factors shown in Table 7, and including random participant and item intercepts. In all models, predictors were standardized (?, ?), and fixed-effect correlations were minimized by centering and residualizing. If the semantic integration predictor was co-linear with any predictors, it was residualized against the other predictors and the semantic integration residuals were entered into the model. After residualization, the fixed-effect correlations with the semantic integration predictor were low (rs $< .27$).

After controlling for other factors, semantic integration did not have a significant effect on D1 duration ($\beta = 0.009$; $t = 0.43$, $p = .66$), N1 duration ($\beta = -0.011$; $t = -0.57$, $p = .61$), or N2 duration ($\beta = -0.015$; $t = -0.58$, $p = .65$). But semantic integration did have an effect on duration at the other three positions: P1, D2, and A1.

Preposition 1 (P1)

After controlling for other factors, higher semantic integration was associated with shorter P1 durations ($\beta = -0.052$; $t = -2.83$, $p < .01$). The average P1 duration was 134ms, and with all else being equal, at the maximum semantic integration rating the model predicted a 127ms preposition while at the minimum rating the model predicted a 142ms preposition. Thus the range in semantic integration in these preambles predicted a 15ms range in P1 duration (11% of its average duration).

Determiner 2 (D2)

After controlling for other factors, higher semantic integration was associated with shorter D2 durations ($\beta = -0.107$; $t = -2.88$, $p < .01$). The average D2 duration was 99ms, and with all else being equal, at the maximum semantic integration rating the model predicted an 89ms D2 while at the minimum rating the model predicted a 111ms D2. Thus the range in semantic integration observed in these preambles predicted a 22ms range in D2 duration (22% of its average duration).

Figure 3: Standardized semantic integration coefficients for each duration model in Experiment 2, with 95% confidence intervals obtained from MCMC sampling. A negative coefficient indicates shorter duration as semantic integration increased. Bars shown in dark gray indicate a significant effect of semantic integration on duration, and larger bars indicate stronger effects. D = Determiner; N = Noun; P = Preposition; A = Adjective; SI = semantic integration.

Adjective 1 (A1)

After controlling for other factors, higher semantic integration was associated with shorter A1 durations ($\beta = -0.082$; $t = -4.33$, $p < .001$). The average A1 duration was 311ms, and with all else being equal, at the maximum semantic integration rating the model predicted a 286ms A1 while at the minimum rating the model predicted a 340ms A1. Thus the range in semantic integration observed in these preambles predicted a 54ms range in A1 duration (17% of its average duration).

Summary

Figure 3 shows the standardized semantic integration coefficient in each of the individual models. D2 showed the largest proportional change in duration as a function of integration, followed by the next largest on A1, and the smallest significant effect on P1. Thus the entire preamble duration was not affected by semantic integration; instead, only the region between N1 and N2 (the head noun and local noun, respectively) was affected, with decreased temporal separation between the head and local noun when they were more integrated. Identical models were used to analyze the data excluding tokens in which participants produced an agreement error. These analyses produced results statistically identical to those including agreement error cases.

Discussion

These findings show that semantic integration, a property that has been hypothesized to affect timing of planning, is reflected in the prosodic timing of an utterance. As predicted by ? (?), the temporal separation of more integrated elements was smaller than the temporal separation of less integrated elements. This effect was most pronounced on D2, but also present for the entire region between the head and local nouns (P1, D2, A1).

It is well documented that upcoming material that is less accessible or less predictable due to frequency effects during lemma/lexeme access, or grammatical encoding properties such as subcategorization preferences, is associated with lengthening. This lengthening can be realized in word durations or by the inclusion of optional words (e.g., ?, ?, ?, ?, ?, ?). Importantly, effects of lexical and grammatical accessibility and predictability cannot explain the pattern observed in Experiment 2, as these properties were either controlled statistically or did not differ in the preambles. Thus, the results of Experiment 2 are consistent with the hypothesis that lengthening may also be observed when subtle conceptual-level factors modulate accessibility of upcoming material (?, ?).

There are at least two possible reasons why the largest proportional change in duration due to semantic integration was observed on D2. First, Solomon and Pearlmutter's (2004) participants were reading preambles aloud, so the eye-voice span may have had an influence. When reading aloud, fixations are approximately two words ahead of the word being articulated (?, ?, ?). This would mean that as speakers in ?'s study were reading the word(s) that determined the semantic integration relationship (A1 and N2), they would have most likely been articulating D2; thus, the most lengthening may be observed in this position. A second possible reason is that determiners are good candidates for lengthening: They tend to appear before content words which vary in many ways and may require additional processing resources, thus determiners may be relatively more likely to adjust duration in response to processing needs (?, ?, ?, ?).

Regardless of why the effects of semantic integration appear largest on D2, these results show that only some of the words within these utterances were affected by semantic integration. It was only the words linking the head and local nouns (N1 and N2) that were affected, with shorter durations observed if the words linked more integrated nouns. Thus, Experiment 2 provides strong evidence in favor of Solomon and Pearlmutter's (?) hypothesis that semantic integration affects the timing of planning of elements within utterances, with more integrated elements being planned more simultaneously (see also ?, ?, ?). The fact that decreased separation of more integrated elements was observed in word durations suggests that the planning time differences that arise through semantic integration at the conceptual level potentially affect all levels of production, even articulation, via the cascading nature of the production process (?, ?).

General Discussion

These findings provide the first direct evidence that semantic integration, a conceptual level property, affects the timing in language production, and thus in turn the prosodic realization of an utterance. Additionally, these results provide more evidence that prosodic separation and timing reflect planning processes in sentence production. ? (?) showed that syntactically and semantically related elements are more likely to appear in the same intonational phrase and hypothesized that this was because speakers are unlikely to place boundaries between elements that are more likely to be planned simultaneously. Experiment 1 showed evidence that speakers were more likely to prosodically separate less-integrated material with prosodic landmarks that are associated with lengthening (2s and 3s in ToBI, ?, ?). In addition, speakers produced evidence of decreased temporal separation of words linking N1 and N2 when elements were tightly integrated versus less integrated. Experiment 2 did not show evidence of acoustically/durationally marked break placement being modulated by semantic integration; however, it does provide evidence of increased temporal separation between less semantically linked elements, which suggests that duration measures may reflect planning time and may capture more subtle properties that intonational breaks may not.

The prosodic structure and timing of an utterance may be affected by multiple factors. Traditionally, it has been argued that prosodic variation and the choice of particular prosodic realization of an utterance are largely driven by the syntactic and/or semantic properties of a speaker's message (e.g., ?, ?); however it has also been noted that prosodic structure does not always correspond to syntactic structure (?, ?), and there is increasing evidence that processing needs of the speaker may be responsible for at least some aspects of prosodic variation (e.g., ?, ?, ?, ?, ?).

The results of Experiments 1 and 2 show that prosodic separation and timing are affected by semantic integration; however, it is not entirely clear whether these prosodic differences are the result of speakers prosodically signaling message-based differences across integration conditions or if they are the result of speaker-internal processing demands that differ across conditions. Pearlmutter and colleagues (?, ?, ?, ?) suggest that semantic integration affects the timing of lexical access, with more integrated elements being accessed and planned for production with more temporal overlap than less integrated elements. Thus, the duration results in Experiments 1 and 2 are consistent with Pearlmutter and colleagues' explanation of semantic integration effects and are likely to reflect the processing differences that integration differences cause during lexical access, provided that the timing differences cascade throughout all levels of production.

In both studies, there were no significant effects of semantic integration on the nouns involved in the integration relationships (N1, N2, N3). Semantic integration is hypothesized to affect the relative timing of lexical access, with more integrated elements being activated closer together in time than less integrated elements (?, ?). Thus semantic integration is hypothesized to affect *when* the nouns in the integration relationship are activated. However, the semantic integration hypothesis does not predict that the activation curve of the integrated elements is affected (see ?, ?, ? for additional details). Thus, the lack of an effect on the nouns in the preambles is not problematic for the semantic integration hypothesis.

Unlike in Watson et al.'s studies (?, ?, ?, ?), the prosodic breaks observed in Experiment 1 that varied as a function of semantic relatedness were not associated with well-formed prosodic constituents, as the majority of breaks observed after PP1 were 2s in ToBI. Given the duration results of Experiments 1 and 2, it is possible that the break differences observed in Experiment 1 were more driven by differences in processing than because speakers use these prosodic breaks to indicate differences in semantic integration. This opens the possibility that break indices that are identified as 2 in ToBI, which are perceptually distinct from break indices of 3 and 4, largely reflect processing-related prosodic variation, whereas the well-formed prosodic constituents indicated by break indices of 3 and 4 in ToBI are mostly message-driven, with additional processing-related influences (e.g., ?, ?, ?). Further work will be necessary to understand how prosody is influenced by processing and message properties, as both are likely to impact the prosodic realization of an utterance (?, ?).

In both experiments, the prosodic patterns roughly map onto subject-verb agreement error patterns in ? (?) and ? (?). In ?, agreement errors were more likely to occur when N2 was plural than when N3 was plural in the early-integrated conditions, but errors were equally likely to be produced whether N2 or N3 was plural in the late-integrated conditions. In Experiment 1, breaks were more likely to follow PP1 in the early-integrated conditions than in the late-integrated conditions, thus prosodically separating N3 from N1 more often in the early-integrated conditions than in the late-integrated conditions. Thus, the local noun that caused the least interference during agreement computation (i.e., the early-integrated N3) was the most likely to be separated from the head noun by prosodic breaks. Both Experiments 1 and 2 also provide evidence from word durations that the temporal separation of tightly integrated elements was reduced compared to the separation of less integrated elements, consistent with predictions derived from agreement error studies examining semantic integration (?, ?, ?). Future research should examine how prosodic timing might influence the domains in which agreement errors occur (and potentially other types of speech errors; see ?, ?, ?).

One concern with these experiments is that the task required a comprehension component; speakers read the preambles prior to completing them as full sentences. Thus, some of the effects observed may have been due to aspects of the comprehension process (for discussion of this concern see ?, ?, ?, ?). Interestingly, there is recent evidence that word duration results in prepared production may be similar to those observed in self-paced reading paradigms used in language comprehension studies (?, ?); however, when participants were producing the sentences they were more likely than when they were reading the sentences to "speed up" (i.e., reduce word duration) during parts of the sentence that were syntactically and semantically linked (e.g., the subject and predicate of a sentence). Thus, while comprehending and producing utterances may rely on similar mechanisms and may show similar word-by-word reading and production times, there are situations in which production differs from comprehension. The semantic integration-related duration changes observed in Experiments 1 and 2 are consistent with the "speed-up" effects observed by ? (?): Durations of words linearly linking related elements were shorter than those linking less related elements.

What these results have shown is that semantic integration, a conceptual-level factor, has an influence at multiple levels of production. ? (?) and ? (?) demonstrated that high integration was associated with increased interference during grammatical encoding resulting in agreement and phrase exchange errors (see ?, ?, ?). Experiment 1 showed that prosodic break production and word durations were affected by semantic integration, with less integrated elements being more likely to be prosodically separated than tightly integrated elements. Experiment 2 provided evidence that semantic integration affects the temporal separation between words at the phonological encoding level. Whether this variation is a direct influence of conceptual factors on a separate prosodic representation that is explicitly planned based on properties of the message, or a result of top-down cascading processing, is still an open question. Even though the source of these effects cannot be determined from these experiments, the results provide stronger support for the hypothesis that semantic factors influence the timing of planning of elements during utterance formulation. Additionally, this work provides a first step as to how researchers can combine prosodic analysis with other methods to address the issues of planning in production.

References

Acknowledgments

We would like to thank Keith Levin and Daryl Velez for ToBI labeling and Mariah Warren for labeling word boundaries. Portions of this work have been presented at the 2009 CUNY Conference on Human Sentence Processing in New York, NY; the 2010 International Workshop on Language Production in Edinburgh, Scotland; and Experimental and Theoretical Advances in Prosody–2 in Montreal, QC. The first author was partially supported by NIH grant 5T32HD055272 (PI: Kay Bock). The third author was partially supported by NSF grant BCS-1023596.