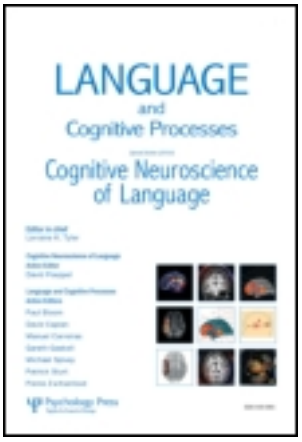

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp20

Intonational phrasing is constrained by meaning, not balance

Mara Breen [a], Duane G. Watson [b] & Edward Gibson [c]

[a] Department of Psychology, University of Massachusetts, Amherst, MA, USA

[b] Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[c] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Available online: 24 May 2011

To cite this article: Mara Breen, Duane G. Watson & Edward Gibson (2011): Intonational phrasing is constrained by meaning, not balance, Language and Cognitive Processes, 26:10, 1532-1562

To link to this article: http://dx.doi.org/10.1080/01690965.2010.508878

LANGUAGE AND COGNITIVE PROCESSES
2011, 26 (10), 1532–1562

Psychology Press
Taylor & Francis Group

Intonational phrasing is constrained by meaning, not balance

Mara Breen[1], Duane G. Watson[2], and Edward Gibson[3]

[1]Department of Psychology, University of Massachusetts, Amherst, MA, USA

[2]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

This paper evaluates two classes of hypotheses about how people prosodically segment utterances: (1) meaning-based proposals, with a focus on Watson and Gibson's (2004) proposal, according to which speakers tend to produce boundaries before and after long constituents; and (2) balancing proposals, according to which speakers tend to produce boundaries at evenly spaced intervals. In order to evaluate these proposals, we elicited naïve speakers' productions of sentences systematically varying in the length of three postverbal constituents: a direct object, an indirect object (a prepositional phrase), and a verb phrase modifier, as in the sentence, *The teacher assigned the*

Correspondence should be addressed to either of the following two authors: Mara Breen, Department of Psychology, University of Massachusetts, 522 Tobin Hall, Amherst, MA 01003, USA. E-mail: mbreen@psych.umass.edu or Edward Gibson, Department of Brain and Cognitive Sciences, MIT 43 Vassar St., Rm. 3035, Cambridge, MA 02139, USA. E-mail: egibson@mit.edu

This material is based upon work supported by the National Science Foundation under Grant No. 0218605, "Intonational boundaries in sentence production and comprehension". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would like to thank Vivek Rao and Jennifer Ford for their help in data coding. We would also like to thank the following people for their comments on earlier presentations of this work: Three anonymous reviewers, Timothy Desmet, Laura Dilley, Fernanda Ferreira, Edward Flemming, Michael Frank, Doug Rohde, Stefanie Shattuck-Hufnagel, Michael Wagner, and the audience at the 2005 Architectures and Mechanisms in Language Processing Conference in Ghent, Belgium. Finally, we would especially like to thank Ev Fedorenko for her extremely helpful and detailed feedback on earlier versions of this paper. Her comments greatly improved the final product.

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2010.508878

chapter (on local history) to the students (of social science) yesterday/before the first midterm exam. Mixed-effects modelling was used to analyse the pattern of prosodic boundaries in these sentences, where boundaries were defined either in terms of acoustic measures (word duration and silence) or following the ToBI (Tones and Break Indices) prosodic annotation scheme. Watson and Gibson's (2004) meaning-based proposal, with the additional constraint that boundary predictions are evaluated with respect to local sentence context rather than the entire sentence, significantly outperformed the balancing alternatives.

Keywords: Prosody; Phrasing; ToBI; Intonation; Syntax; Language production.

In language production, speakers slow down, or use a filler like "uh" or "um", when they have not fully planned what they are going to say or when they cannot access a particular word or expression (e.g., Arnold & Tanenhaus, 2010; Clark & Wasow, 1998; Fox, Tree, & Clark, 1997). Furthermore, even when speakers have planned what they want to say, they will tend to break up long utterances into shorter segments (e.g., Ferreira, 2007; Gee & Grosjean, 1983; Watson & Gibson, 2004). For example, consider sentence (1):

(1) The professor assigned the chapter on local history to the students of social science yesterday.

Even if speakers have fully planned an utterance like (1), they will typically segment such an utterance by placing intonational boundaries between certain words and phrases. In this paper, we investigate the constraints on intonational boundary production.

Characteristic acoustic cues mark the location of intonational boundaries. Speakers typically mark boundaries with increased duration of preboundary words (Ferreira, 1993; Lehiste, Olive, & Streeter, 1976; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Schafer, Speer, Warren, & White, 2000; Selkirk, 1984; Snedeker & Trueswell, 2003; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992) and/or silence (Cooper & Paccia-Copper, 1980; Klatt, 1975; Lehiste, 1973). In addition, speakers often raise or lower their pitch at the end of intonational phrases (Pierrehumbert, 1980; Streeter, 1978).

This paper evaluates two classes of hypotheses about how speakers prosodically segment utterances. According to one general class of proposals—meaning-based proposals—the acoustic properties of the linguistic signal (in this case, prosodic grouping) reflect the syntactic and semantic interpretation of an utterance (Cooper & Paccia-Cooper, 1980; Ferreira, 1988; Gee & Grosjean, 1983; Watson & Gibson, 2004). In particular, boundaries are hypothesised to reflect the grouping of words/phrases into syntactically and semantically coherent units. The empirical predictions of different proposals in this class are similar, but Watson and Gibson's (2004, henceforth W&G) proposal is conceptually the simplest. We therefore focus on W&G's proposal

as a representative of the class of meaning-based proposals. Although the initial empirical evaluations presented by W&G in support of their proposal were promising, there are methodological limitations of W&G's original study that make it difficult to draw strong conclusions from that work. Therefore, the first goal of the current paper is to provide a more rigorous evaluation of W&G's proposal.

According to an alternative class of proposals for prosodic segmentation (balancing-based proposals), boundaries occur at regular intervals, resulting in prosodic units of roughly equal length (Augurzky, 2008; Cooper & Paccia-Cooper, 1980; Fodor, 1998; Gee & Grosjean, 1983; Grosjean, Grosjean, & Lane, 1979). No concrete versions of balancing-based proposals have been previously proposed or evaluated independently of meaning-based proposals. The second goal of the current paper is therefore to propose some potential balancing algorithms, and determine how much variance in boundary placement these algorithms explain, and whether they capture variance beyond W&G's meaning-based model.[1]

In the following sections, we (1) describe W&G's meaning-based proposal; and (2) present three possible versions of balancing. We then proceed to evaluate the predictions of these proposals using a production experiment.

A MEANING-BASED PROPOSAL: RECOVERY AND PLANNING

It has long been observed that an ambiguous sentence can be produced with different intonational segmentation corresponding to different meanings. For example, Price et al. (1991) had speakers produce ambiguous sentences like those in (2) following a disambiguating context, and showed that speakers consistently place boundaries in different places depending on the intended meaning.

(2) When you learn gradually you worry more.
a. When you learn gradually | you worry more.
b. When you learn | gradually you worry more.

When speakers intend the meaning whereby *gradually* is modifying the verb *learn*, they tend to place a boundary after *gradually*, as in (2a). Alternatively, if speakers intend the meaning whereby *gradually* is modifying the clause *you*

[1] Another dimension of proposals for prosodic segmentation—audience-design vs. speaker-internal constraints—is not investigated in the current study. The results are generally consistent with either kind of proposal: one where speakers segment their utterances in order to be maximally comprehensible to their listeners (Bell, 1984; Clark & Wilkes-Gibbs, 1986), or one where speakers segment their utterances because of speaker-internal constraints (e.g., Ferreira, 1993, 2007; Gee & Grosjean, 1983;).

worry more, they tend to place a boundary after *learn*, as in (2b). More generally, several production studies have demonstrated that speakers tend to use boundaries to signal that upcoming words do not lexically depend on the immediately preceding words (Albritton, McKoon, & Ratcliff, 1996; Kraljic & Brennan, 2005; Schafer et al., 2000; Snedeker & Trueswell, 2003). Correspondingly, several researchers have presented proposals that attempt to account for this and other kinds of meaning-based prosodic boundary effects (Cooper & Paccia-Cooper, 1980; Ferreira, 1988; Gee & Grosjean, 1983; W& G, 2004).

According to meaning-based proposals, the presence of longer syntactic constituents increases the probability of an intonational boundary before and/or after the long constituent. Gee and Grosjean's, Cooper and Paccia-Cooper's and Ferreira's proposals, although all successfully predict boundary placement in their respective data sets, are complex, requiring many steps to get from a syntactic parse to a candidate intonational phrasing. W&G proposed a simpler version of a meaning-based proposal, which included two components: *recovery* and *planning*. First, W&G observed that the size of the most recently completed constituent was a good predictor of a boundary, regardless of the constituent's hierarchical position in a syntactic tree of the sentence. The recovery component was thus defined in terms of the size of the most recently completed constituent (i.e., material to the left of the boundary), where a constituent is complete if it has no obligatory rightward dependents. W&G further observed that a boundary is more likely before the production of a longer syntactic constituent.[2] The planning component was therefore defined in terms of the size of upcoming material (i.e., material to the right of the boundary). Furthermore, W&G's model included a constraint that words and constituents that rely on one another for meaning should be produced in the same intonational phrase (cf. the sense unit condition of Selkirk, 1984). Empirical evidence for the concurrent planning of semantically integrated material comes from Watson, Breen, and Gibson (2006), who demonstrated that (1) arguments and their heads are more likely to occur in the same prosodic phrase than adjuncts and their heads; and (2) obligatory arguments are more likely to occur in the same prosodic phrase as their heads than nonobligatory arguments (see also Solomon & Pearlmutter, 2004). Finally, following empirical observations of Gee and Grosjean (1983), W&G constrained boundaries so that they could not be produced within

[2] W&G discuss each of these factors in terms of speaker-internal constraints (as opposed to audience-design). In particular, they discuss the recovery component of their proposal as due to a 'refractory period' in which the speaker must recover from the expenditure of resources involved in producing utterances, and they discuss the planning component as reflecting the time that is needed for planning longer upcoming constituents. But both of these components can be conceived of in terms of audience-design. W&G did not evaluate this aspect of their proposal.

phonological phrases, where a phonological phrase is defined as a head noun or verb and all the material that comes between that head and the preceding one, including function words, prenominal adjectives (e.g., *green* in *green shirt*), and preverbal adverbs (e.g., *gradually* in *gradually learn*) (Nespor & Vogel, 1986).

W&G argued that the main advantage of their model over previous models was its relative conceptual simplicity. W&G evaluated the predictions of their recovery–planning proposal against those of other meaning-based proposals in two production experiments and observed that the recovery–planning model performed similarly to the others in the first experiment, and better than the others in the second. However, there were several methodological limitations in W&G's empirical evaluation of their proposal. First, recovery and planning were essentially post-hoc formulations, tested on the data they were designed to explain. W&G did test the predictions on an independent data set in Experiment 2, but only on one variant of a particular syntactic construction. Second, determining a model's efficacy using a single correlation coefficient averaged over sentences, like that reported by W&G, provides only a coarse-grained evaluation (Ferreira, 2007). And third, W&G used relative clauses to increase constituent length in the majority of sentences in Experiment 1, as shown in (3). The recovery–planning model predicts the highest probability of boundaries at locations [a] and [b], due to the length of material preceding (for [b]) and following (for [a]) these locations. Indeed, these locations had the highest probability of boundaries in speakers' productions. However, these locations also mark the beginning and end of a relative clause. Relative clauses, when interpreted nonrestrictively, are often produced in separate intonational phrases for reasons having to do with discourse organisation, rather than factors like recovery and planning (Dehé, 2002; Selkirk, 1984). Consistent with this idea, Experiment 3 from W&G (2004) demonstrated that, when discourse supported a nonrestrictive reading of a relative clause, speakers were significantly more likely to place boundaries after the head noun ([a] below).

(3) The judge [a] who the reporter ignored [b] fired the secretary.

The current study addresses these limitations. First, the current study manipulates recovery and planning demands independently in the same set of materials to yield predictions for 40 test positions across eight experimental conditions, as presented in (4). This study thus provides the first rigorous evaluation of W&G's proposal using a data set that was not used to formulate the model. Second, mixed-effects regression models are used in the analyses, allowing for an evaluation of the simultaneous influence of multiple factors on the variables of interest. And third, the materials in the current experiment do not use relative clauses in order to manipulate constituent length, with the consequence that boundary production is less likely to be driven by

discourse-level factors. In addition to addressing the weaknesses of W&G's empirical evaluation of their proposal, the design of the materials in the current experiment allows us to compare the predictions of W&G's recovery–planning model against those made by three alternative, not meaning-based, proposals, as discussed below.

(4) a. Short direct object (DO), Short indirect object (IO), Short modifier (Mod)
The professor assigned the chapter to the students yesterday.
b. Long DO, Short IO, Short Mod
The professor assigned the chapter on local history to the students yesterday.
c. Short DO, Long IO, Short Mod
The professor assigned the chapter to the students of social science yesterday.
d. Long DO, Long IO, Short Mod
The professor assigned the chapter on local history to the students of social science yesterday.
e. Short DO, Short IO, Long Mod
The professor assigned the chapter to the students after the first midterm exam.
f. Long DO, Short IO, Long Mod
The professor assigned the chapter on local history to the students after the first midterm exam.
g. Short DO, Long IO, Long Mod
The professor assigned the chapter to the students of social science after the first midterm exam.
h. Long DO, Long IO, Long Mod
The professor assigned the chapter on local history to the students of social science after the first midterm exam.

THRESHOLDED RECOVERY–PLANNING

In the original formulation of W&G's recovery–planning model, there is no limit on how much preceding or following sentence material can count towards the recovery and planning weights. This version of the model may not be psychologically plausible, as prior work demonstrates that the production system is incremental, such that there are limits on how much sentence material is planned prior to the initiation of speaking (Allum & Wheeldon, 2007; Brown-Schmidt & Konopka, 2008; Brown-Schmidt & Tanenhaus, 2006; Ferreira & Swets, 2002; Garrett, 1980). To address this weakness of the recovery–planning model, we propose the *thresholded recovery–planning* model. Under this model, there is a maximum threshold on the amount of material to be considered in recovery and planning. The thresholded recovery–planning model thus captures the incrementality of production, while taking into account the meaning of the sentence. We tested thresholds between one and four phonological phrases, and a threshold of

two resulted in the best fit of the current data. We therefore present the results from the version of the model with this threshold.

BALANCING-BASED PROPOSALS

According to early proposals about boundary placement, speakers split their utterances into two equal-length segments, resulting in roughly "balanced" phrasing (Bachenko & Fitzpatrick, 1990; Cooper & Paccia-Cooper, 1980; Grosjean et al., 1979; Wang & Hirschberg, 1992). Although these models were successful at predicting boundaries in the corpora on which they were tested, the models also included syntactic constraints on boundary placement, so it is unclear which constraints led to their success. In fact, Gee and Grosjean (1983) concluded that the balanced phrasing that they observed was a by-product of syntactic grouping and disallowing boundaries within phonological phrases. Here, we propose three different ways of formalising balancing constraints, so that they can be evaluated independently from meaning-based grouping constraints.

1. Fixed phrase length

According to the first balancing hypothesis we consider—the *fixed phrase length* hypothesis—the speaker produces prosodic phrases of a fixed length. In order to work out the predictions of this hypothesis, we need to specify (a) a metric for measuring length; and (b) a value for the length of a prosodic phrase. Plausible candidates for measuring length include syllables, words, and phonological phrases. Results from initial analyses of the current data suggested that it is unlikely that boundary placement is based on number of syllables or number of words. In particular, in the current data set, the first noun phrase of the sentence (e.g., *The professor* in (4)) could range in length from two to five syllables. If an increase in the number of syllables leads to an increase in boundary probability, we would expect a correlation between the length (in syllables) of the sentence-initial noun phrase and the probability of a boundary. However, there was no increase in boundary probability as the number of syllables increased, whether boundaries were measured in terms of ToBI (Tones and Break Indices; Silverman et al., 1992) boundary labels ($r = -.015$) or postword silence ($r = -.007$) (see Section "Data Coding" below). Similarly, words are also unlikely to be the right grain of measure, as speakers rarely place a boundary within a phonological phrase in fluent utterances (Gee & Grosjean, 1983; Nespor & Vogel, 1986; W&G, 2004; Watson et al., 2006). In the current study, for example, speakers only placed boundaries inside phonological phrases 16 times out of a possible 3,964 instances (0.4%) across the entire data set. These findings suggest that

if the language production system has a fixed phrase length, then a plausible unit of length is the phonological phrase.

In order to empirically determine the optimal fixed phrase length for our data set, we tested all lengths from one to five phonological phrases with respect to people's productions of the materials described above. The fixed phrase length model which best accounted for boundary placement in our materials was one with a fixed length of two phonological phrases. Therefore, in what follows we will discuss the predictions of a two-phrase fixed phrase length hypothesis.

Note that a fixed phrase length in production will often predict boundaries to occur immediately following a verb, depending on the length of the material before the verb (e.g., it will always predict a boundary after *assigned* in (4)). Because previous research has demonstrated that speakers rarely place boundaries between verbs and their direct objects (Ferreira, 1991; Nespor & Vogel, 1986; W&G, 2004; Watson et al., 2006), a fixed phrase length metric is unlikely to fit the data well. We therefore also evaluate a modified version of the fixed phrase length proposal, *hybrid fixed length*, which prohibits boundaries between verbal heads and their obligatory arguments (i.e., after *assigned* in (4)). Hybrid fixed length, like fixed phrase length, predicts the highest probability of a boundary after every two phonological phrases, unless the second phonological phrase ends on a verbal head when the verb has an obligatory argument following it (always the case in our materials), in which case the next boundary location is predicted to be at the following phonological phrase boundary.

2. Prior boundary

Several prior studies have hypothesised that a speaker's prior boundary placement may influence subsequent boundary placement (Sanders & Taylor, 1995; Wang & Hirschberg, 1992), but this hypothesis has not been properly evaluated. The second balancing hypothesis that we consider, therefore, is the *prior-boundary* hypothesis. The prior-boundary hypothesis is like the fixed phrase length hypothesis in that the likelihood of a boundary is calculated based on the speaker having just produced an intonational boundary. But rather than there being a certainty of producing the next boundary a fixed number of phonological phrases later, the probability of a boundary under the prior-boundary hypothesis increases linearly with the number of phonological phrases from the last boundary location.

3. Even spacing

A third way that the language production system may have adapted in order to result in balanced output relies not only on knowing that an intonational boundary has just been produced, but also on knowing the length of the

upcoming material (similar to the planning component of the recovery–planning model, but not taking into consideration the syntactic/semantic properties of the material, only its length). Under this *even-spacing* hypothesis, the producer places boundaries at evenly spaced intervals, dividing the utterance into phrases of the same size (e.g., Grosjean et al., 1979). This proposal differs from the fixed phrase length hypothesis when the threshold T does not divide evenly into the length of the utterance to be produced, L. In such a case, the even-spacing algorithm finds the largest factor F of L that is less than T, and produces L/F segments. When there are no factors of L less than T and greater than 1, then the even-spacing proposal reverts to using the threshold T to divide the utterance, placing "leftover" material at either the beginning or end of the utterance, to be discussed below.

For example, if the threshold is five units ($T = 5$), and the length of the utterance to be produced is six units ($L = 6$), the even-spacing hypothesis produces two segments of length three ($F = 3$), rather than a segment of length five followed by a segment of length one, as predicted by the fixed phrase length hypothesis with a fixed length of five. This proposal is therefore similar to the fixed phrase length hypothesis, with two important differences: (a) the speaker must have access to knowledge about the entire sentence; and (b) the absolute size of the resulting prosodic phrases can vary with the length of the sentence.

Several parameters were considered in order to cover the space of possible even-spacing algorithms. These include: (a) the size of the threshold T; (b) what to do when there is no factor of L less than T, i.e., when the utterance cannot be evenly segmented; and (c) a specification of how even spacing interacts with other known constraints on phrasing, such as boundaries not being possible between verbs and their obligatory arguments. The values that we considered for these parameters are listed in (5):

(5) a. Threshold value: We tested models where the threshold T varied from 2 to 6 phonological phrases.

b. Leftover-first vs. leftover-last: We tested models where, when the length of the projected utterance L doesn't divide evenly, the leftover section is produced early (first) or later (last). For example, if $T = 3$, and $L = 5$, then leftover-first constraint leads to a production of 3, 2; whereas leftover-last leads to a production of 2, 3.

c. With/without blocking boundaries between verbs and their obligatory arguments: We tested models in which boundaries are either allowed or disallowed between verbs and their obligatory arguments. In cases where this constraint conflicted with others, we tested an algorithm where the predicted boundary would occur at the phonological phrase boundary preceding the verb.

All possible combinations of values for these three parameters resulted in 14 possible even-spacing models. (It is only 14 possible even-spacing models and not 5*2*2 = 20 models because many of these versions are identical to one another.) One version was clearly the best at predicting boundary location in the corpus that was generated from our experiment. This model resulted in a log likelihood of 1,262, which was much higher than the other models, which scored between 1,163 and 1,203 (see Section "Results" for more details regarding the interpretation of this kind of analysis). This model assumes (a) a threshold of two phonological phrases; (b) that leftover material gets produced last; and (c) that boundaries are disallowed between verbs and their obligatory arguments. In order to simplify the presentation, we will restrict the discussion of our evaluation of even-spacing algorithms to this model.

Note that the even-spacing hypothesis requires the production system to have a large amount of look-ahead. That is, in order to determine where to place intonational boundaries in order to divide an utterance evenly, the speaker needs know the approximate length of the complete utterance. As stated above, this amount of look-ahead is likely not available to the producer, so an even-spacing model of this kind may not be psychologically plausible.

PREDICTIONS

There are six proposed algorithms to be evaluated: (1) W&G's recovery–planning proposal; (2) the thresholded recovery–planning proposal; (3) the fixed phrase length hypothesis; (4) the hybrid fixed-length hypothesis; (5) the prior-boundary hypothesis; and (6) the even-spacing hypothesis. In order to test the recovery–planning model and the thresholded recovery–planning model, we simultaneously manipulated the length of material both preceding and following a possible boundary location. In this way, we could see if the presence of more material that is semantically related to a head word before or after a possible boundary location would lead to a greater probability of boundaries at that location. In order to test the predictions of the fixed phrase length, hybrid fixed-length, and even-spacing hypotheses, we included sentences of varying lengths. Finally, in order to test the prior-boundary hypothesis, we manipulated the length of three sentence constituents to increase the number of possible boundary locations.

The predictions of each of the six proposals are presented in Table 1. For the recovery–planning model, the boundary weight is the sum of a recovery weight and a planning weight. The recovery weight is the number of phonological phrases of the largest most recently completed constituent. The planning weight is the number of phonological phrases of the largest complete upcoming constituent. For the thresholded recovery–planning

TABLE 1

Predictions of five of the six models of boundary prediction at each phonological phrase boundary for a sample item. Predictions for the prior-boundary model cannot be presented in this format, because this model makes its predictions on a trial-by-trial basis, so that there is no meaningful average response at each location across trials. Numbers correspond to the number of phonological phrases counted by each metric; 'X' indicates that a boundary is disallowed according to the given hypothesis

4a(4e)	The professor	assigned	the chapter	to the students	yesterday (before the first midterm exam)
Fixed phrase length	1	2	1	2	
Hybrid fixed length	1	X	2	1	
Even spacing	2	X	2	1(2)	
Recovery	1	0	1	1	
Planning	3	0	1	1(3)	
tRecovery	1	0	1	1	
tPlanning	2	0	1	1(2)	

4b(4f)	The professor	assigned	the chapter	on local history	to the students	yesterday (before the first midterm exam)
Fixed phrase length	1	2	1	2	1	
Hybrid fixed length	1	X	2	1	2	
Even spacing	2	X	2	2(1)	1(2)	
Recovery	1	0	1	2	1	
Planning	4	0	1	1	1(3)	
tRecovery	1	0	1	2	1	
tPlanning	2	0	1	1	1(2)	

TABLE 1 (*Continued*)

4c(4g)	The professor	assigned	the chapter		to the students	of social science	yesterday (before the first midterm exam)
Fixed phrase length	1	2	1		2	1	
Hybrid fixed length	1	X	2		1	2	
Even spacing	2	X	2		2(1)	1(2)	
Recovery	1	0	1		1	2	
Planning	4	0	2		1	1(3)	
tRecovery	1	0	1		1	2	
tPlanning	2	0	2		1	1(2)	
4d(4h)	The professor	assigned	the chapter	on local history	to the students	of social science	yesterday (before the first midterm exam)
Fixed phrase length	1	2	1	2	1	2	
Hybrid fixed length	1	X	2	1	2	1	
Even spacing	2	X	2	1(2)	2(1)	1(2)	
Recovery	1	0	1	2	1	2	
Planning	5	0	1	2	1	1(3)	
tRecovery	1	0	1	2	1	2	
tPlanning	2	0	1	2	1	1(2)	

model, the boundary weight is again the sum of a recovery weight and a planning weight, but both of these weights are thresholded at two, resulting in a maximum total weight of four. For fixed phrase length, the boundary weight is the number of phonological phrases that intervene between a potential boundary location and the most recent predicted boundary, as long as that number is not larger than two. For hybrid fixed length, the boundary weight is the number of phonological phrases that intervene between a potential boundary location and the most recent predicted boundary, provided that (a) the current point is not between a verbal head and its obligatory argument; and (b) the weight is not larger than two. For the prior-boundary hypothesis (whose predictions cannot be presented in Table 1 because the weights are computed on a trial-by-trial basis), the boundary weight is the number of phonological phrases back to the last boundary produced in the sentence. For the even-spacing hypothesis, the weight is the number of phonological phrases that intervene between a potential boundary location and the most recent predicted boundary, provided that (a) the weight is not larger than two; (b) if there is an odd number of phrases, the leftover phrase is grouped with the material after the boundary; and (c) the current point is not between a verbal head and its obligatory argument.

METHOD

Participants

Forty-eight native English speakers from the MIT community participated in the study for $10.00 each. Participants were run in pairs, and each member of the pair was randomly assigned the role of speaker or listener. Data from 3 of the 24 pairs of subjects could not be used due to poor recording quality. Productions from 18 of the remaining 21 pairs were coded for intonational boundaries (using the ToBI annotation scheme) by a coder who was unaware of any theoretical predictions for the materials. Productions from all 21 successfully recorded pairs were analysed for their word durations and the duration of silence following each word.

Materials and design

Length of the direct object, length of the indirect object, and length of the modifier were manipulated in a $2 \times 2 \times 2$ design, as shown in (4), to create 32 items. The long-direct-object and long-indirect-object conditions were created by adding a modifier phrase or nonobligatory argument phrase to the object noun phrase (e.g., the chapter *on local history,* the bouquet *of thirty roses*, the turkeys *with homemade stuffing*). All direct objects and indirect objects in the short conditions had three syllables, while the long

conditions had seven or eight syllables. The short modifiers were temporal modifiers (in 23 items) or adverbs (nine items) comprised of one or two words (two to four syllables total), but always only one phonological phrase (e.g., *yesterday, last night, on Sunday*). The long modifiers were always temporal modifiers containing five words, which were comprised of two or more phonological phrases.

Eight experimental lists were created, following a Latin Square design, such that each participant saw only one version of each item. The order of trials in each list was randomised separately for each participant. Experimental items were randomly interspersed with 44 fillers, which were comprised of items from two other unrelated experiments, with different syntactic structures. The full set of experimental items can be found in Appendix 1.

Procedure

In order to elicit as naturalistic productions as possible, while still maintaining control over the words produced, we used a two-participant reading task. In this method, a speaker reads a sentence aloud for a listener, who must answer a comprehension question about the sentence (Hirotani, 2007; W&G, 2004; Watson et al., 2006). The benefits of this task are as follows: (a) it allows tight control over the material produced; and (b) speakers produce more natural-sounding utterances when they know their partner must comprehend what they say. A possible drawback of this method is that a speaker's production of a read sentence may not match their spontaneous production of a similar structure. However, several studies have noted strong similarity between boundary production in read and spontaneous utterances (e.g., Breen, Dilley, Kraemer, & Gibson, 2010; Ferreira, 1991; Hirschberg, 1995). Nevertheless, evaluating the proposals discussed here using more naturalistic productions would be a powerful extension of this work.

The experiment was conducted using Linger, a software platform for language processing experiments.[3] Two participants—a speaker and a listener—sat at computers in the same room such that neither could see the other's screen. The "speakers" were instructed that they would be producing sentences for their partners (the "listeners"), and that the listeners would be required to answer a comprehension question about each sentence immediately after it was produced. Each trial began with the speaker being presented with a sentence on the computer screen to read silently until she/he understood it. The speaker then answered a multiple-choice content question

[3] Linger was written by Doug Rohde, and can be downloaded at: http://tedlab.mit.edu/ ~dr/Linger/

about the sentence, to ensure understanding. If the speaker answered correctly, she/he proceeded to produce the sentence out loud. If the speaker answered incorrectly, she/he was given another chance to read the sentence, and to answer a different question about it. The speaker always produced the sentence after the second question whether or not she/he got the second question right.

The listener sat at another computer, and saw a blank screen while the speaker went through the procedure described above. After the speaker produced a sentence out loud for the listener, the listener would press the space bar on his/her computer. A multiple-choice question about the content of the sentence just heard would then appear. Listeners were provided feedback when they answered a question incorrectly.

Data coding

Trials where an independent coder identified a disfluency in the production were excluded from analysis, accounting for 4.6% of the data. Trials where either (a) the speaker answered both comprehension questions incorrectly, or (b) the listener answered his/her comprehension question incorrectly accounted for 3.2% of the data and were also excluded.

Each sentence was recorded digitally, and analysed using the PRAAT program (Boersma & Weenink, 2006). We coded boundaries in two ways: (1) based on the ToBI prosodic annotation conventions (Silverman et al., 1992); and (2) based on acoustic measures.

Each production was coded for intonational boundaries by an expert coder using a subset of the ToBI conventions. The coder—who was not an author of the paper—was blind to the predictions of the different proposals. The coder indicated the strength of perceptual disjuncture after each word using the following standard break indices: 4—major intonational phrase boundary; 3—minor phrase boundary; 0, 1, 2—no phrase boundary. We treated all break indices of "0," "1," or "2" as nonboundaries, and all break indices of "3" or "4" as boundaries. Using these criteria, boundaries occurred at 64% of the phonological phrase boundaries.[4]

Perceptual boundary identification is a difficult task, and even expert coders are not in perfect agreement about the presence of boundaries (Breen et al., 2010; Dilley, Breen, Gibson, Bolivar, & Kraemer, 2006; Pitrelli, Beckman, Hirschberg, 1994). In contrast, acoustic measures are objective,

[4] Analyses comparing the recovery-planning model to the fixed phrase length, hybrid length, and prior-boundary models were also performed on the data in which only "4"s were counted as boundaries. These analyses revealed the same patterns as the analyses presented here, but analyses based on 4's alone accounted for significantly less variance than analyses based on the data in which "3"s and "4"s were counted as boundaries, across models. Therefore, we report only the results from the analyses where both "3's" and "4's" are included.

and consequently can be semi-automatically extracted. It has been hypothesised that ToBI coders typically use the acoustic cues of preboundary word duration and postword silence in order to code boundaries (Beckman & Ayers Elam, 1997; Wightman et al., 1992). We therefore also measured preboundary word duration and postword silence in our corpus, as follows: The data set was divided among three additional coders, who were all blind to the hypotheses. Coders observed each trial's spectrogram in PRAAT, and located word boundaries and silence using the guidelines for segment identification presented in Turk, Nakai, and Sugahara (2006). From the word- and silence-aligned transcripts of the productions the coders extracted (a) the duration of each word that preceded a phonological phrase boundary; and (b) the duration of any silence that followed a phonological phrase boundary. The sum of these two measures determined the acoustic measure of the relative strength of a boundary. If ToBI coders indeed use preboundary word duration and postword silence to code boundaries, a strong correlation between each of these measures and boundary presence, as measured by ToBI labels should be observed. Indeed, we observed a strong correlation both between preboundary word duration and ToBI boundary presence ($r = .57$, $N = 2{,}961$, $p < .001$), and between postword silence and ToBI boundary presence ($r = .62$, $N = 2{,}961$, $p < .001$). (It is worth noting that these strong correlations were present despite the fact that each participant produced only one version of each item, suggesting that these acoustic cues are used in a similar way across participants.) These correlations support the idea that preboundary word duration and postword silence strongly contribute to a percept of a boundary.

RESULTS

Boundaries as acoustic measures

Table 2 presents the sum of the word duration and silence at each phonological phrase boundary, averaged across participants for each phonological phrase boundary. In a series of mixed-effects models (cf. Jaeger, 2008), the weights of each of five of the six boundary prediction models in Table 1 were used to predict these duration values. In each analysis, the boundary prediction model being tested was included as a fixed effect, and subjects and items were included as random effects. The modelling results are presented in Table 3. The best-fitting model is determined in two ways: by the size of the log-likelihoods, such that better model fit is indicated by a higher log-likelihood; and by the strength of the correlation between predicted and observed values.

Comparisons of the log-likelihoods indicate that the thresholded recovery–planning model provides the best fit of the six proposals being

TABLE 2
(a) Probability of boundary production as determined by ToBI labels; and (b) the mean summed word duration and silence (in ms) at each phonological phrase boundary for a sample item

(4a)	The professor	assigned	the chapter	to the students	yesterday
Average probability of a boundary	.41	.02	.46	.59	
Word duration + silence (ms)	556	389	543	507	

(4b)	The professor	assigned	the chapter	on local history	to the students	yesterday
Average probability of a boundary	.41	.02	.40	.81	.39	
Word duration + silence (ms)	557	420	538	616	493	

(4c)	The professor	assigned	the chapter	to the students	of social science	yesterday
Average probability of a boundary	.39	.02	.49	.43	.71	
Word duration + silence (ms)	560	404	575	510	529	

(4d)	The professor	assigned	the chapter	on local history	to the students	of social science	yesterday
Average. probability of a boundary	.49	.04	.41	.84	.45	.77	
Word duration + silence (ms)	551	401	526	632	530	542	

TABLE 2 (*Continued*)

(4e)	The professor	assigned	the chapter		to the students		before the first midterm exam
Average probability of a boundary	.49	.01	.45		.74		
Word duration + silence (ms)	545	414	549		555		

(4f)	The professor	assigned	the chapter	on local history	to the students		before the first midterm exam
Average probability of a boundary	.53	.01	.34	.78	.70		
Word duration + silence (ms)	575	394	511	616	554		

(4g)	The professor	assigned	the chapter		to the students	of social science	before the first midterm exam
Average. probability of a boundary	.41	.02	.46		.38	.87	
Word duration + silence (ms)	551	381	548		520	602	

(4h)	The professor	assigned	the chapter	on local history	to the students	of social science	before the first midterm exam
Average probability of a boundary	.36	.20	.34	.81	.41	.86	
Word duration + silence (ms)	550	406	525	641	516	597	

TABLE 3
Summary of the fixed effects in five single-factor mixed-effects models and one two-factor mixed-effects models, all predicting phrase-final word duration and silence ($N = 2{,}961$). There is no model corresponding to the prior-boundary model because computation of this model requires a binary boundary distinction, which is not possible with acoustic data

Fixed effects	*Coefficient*	*SE*	*T*	*p*	*log-likelihood*	r^2
Intercept	0.6	0.02	31.89	<.0001	1,197	.02
Fixed phrase length	−0.05	0.01	−8.42	<.0001		
Intercept	0.46	0.02	26.62	<.0001	1,262	.06
Hybrid fixed length	0.06	0.004	14.43	<.0001		
Intercept	0.47	0.02	27.92	<.0001	1,262	.04
Even-spacing	0.08	0.006	14.38	<.0001		
Intercept	0.44	0.02	25.63	<.0001	1,354	.10
Recovery-planning	0.03	0.002	20.36	<.0001		
Intercept	0.41	0.02	23.57	<.0001	1,435	.17
tRecovery-planning	0.05	0.002	24.55	<.0001		
Intercept	0.41	0.02	23.50	<.0001	1,432	.17
tRecovery-planning	0.05	0.002	19.28	<.0001		
Even-spacing	0.009	0.007	1.37	.18		

compared. In the complementary correlational analyses, it was also observed that the correlation between predicted and observed values is higher for the thresholded recovery–planning model than for the fixed phrase length model ($z = 10.9$, $p < .0001$), the hybrid fixed-length model ($z = 7.17$, $p < .0001$), the even-spacing model ($z = 8.91$, $p < .0001$), or the recovery–planning model ($z = 4.23$, $p < .0001$). The predicted values of the thresholded recovery–planning model are plotted against the observed values in Figure 1.

Boundaries as ToBI labels

Table 2 also presents the probability of a boundary, as coded in ToBI, averaged across participants for each phonological phrase boundary. The weights of each boundary prediction model in Table 1 were regressed against the presence of a boundary in a series of linear mixed-effects logit models. The probability of a boundary was predicted by each model as a fixed effect, with subjects and items as random effects. The model results are summarised in Table 4. A comparison of the log-likelihoods indicates that the thresholded recovery–planning model is the best predictor of boundary production. In the complementary correlational analyses, it was also observed that the correlation between predicted and observed values is higher for the thresholded recovery–planning model than for the fixed phrase length model ($z = 7.52$, $p < .0001$), the hybrid fixed-length model ($z = 6.96$, $p < .0001$), the even-spacing model ($z = 8.15$, $p < .0001$), the prior-boundary model ($z = 4.37$, $p < .0001$), or the recovery–planning model ($z = 4.24$, $p < .0001$).

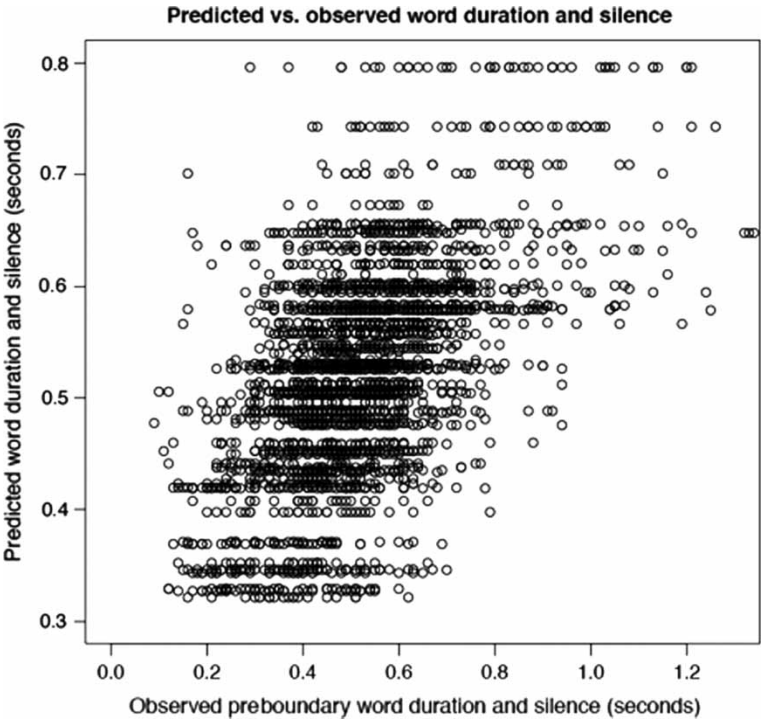

Figure 1. Scatterplot of the relationship between the predicted sum of phrase-final word duration and postword silence computed by the thresholded recovery-planning model in Table 3, and the actual values of word duration and silence. Note that the bulk of the points fall on the line where $x = y$.

Figure 2 presents the relationship between the boundaries predicted by each of the models in Table 4 and the ToBI boundary labels. The x-axis of each panel represents the ToBI boundary data, in which every phonological phrase boundary was labelled with a 0 (no boundary) or 1 (boundary). The y-axes indicate the boundary prediction of each model, defined in terms of a log-odds ratio. Specifically, for a given model, for each phonological phrase boundary, we computed a predicted probability of a boundary. These probabilities were then transformed into log-odds ratios. The box-and-whisker plots represent the distribution of these ratios for phonological phrases that were labelled boundaries (1's), and those that were not (0's). The boxes represent the inter-quartile range of the distribution, while the whiskers represent the highest and lowest quartiles. Better fit of the ToBI boundary data is indicated by a larger separation between the boxes. In the sixth panel, which presents the relationship between the predictions of the thresholded recovery–planning model and the ToBI labels, the

TABLE 4
Summary of the fixed effects in the six mixed-effects single-factor logit models and one mixed-effects two-factor logit models, all predicting intonational boundaries ($N =$ 2,961). Boundaries are determined by ToBI labels

Fixed effects	*Coefficient*	*SE*	*T*	*p*	*log-likelihood*	r^2
Intercept	−2.16	0.21	−10.2	<.0001	−1479	.09
Fixed phrase length	0.57	0.09	6.23	<.0001		
Intercept	−1.97	0.18	−10.8	<.0001	−1464	.1
Hybrid fixed length	0.54	0.07	8.1	<.0001		
Intercept	−1.50	0.17	−8.89	<.0001	−1493	.09
Even-spacing	0.31	0.09	3.29	<.001		
Intercept	−2.73	0.24	−11.32	<.0001	−1397	.14
Prior-boundary	0.53	0.04	13.92	<.0001		
Intercept	−2.51	0.2	−12.88	<.0001	−1391	.15
Recovery-planning	0.4	0.03	13.88	<.0001		
Intercept	−5.09	0.3	−16.96	<.0001	−1179	.22
tRecovery-planning	1.43	0.08	18.73	<.0001		
Intercept	−4.76	0.29	−16.21	<.0001	−1137	.26
tRecovery-planning	1.59	0.08	21.19	<.0001		
Even-spacing	−1.10	0.17	−9.42			

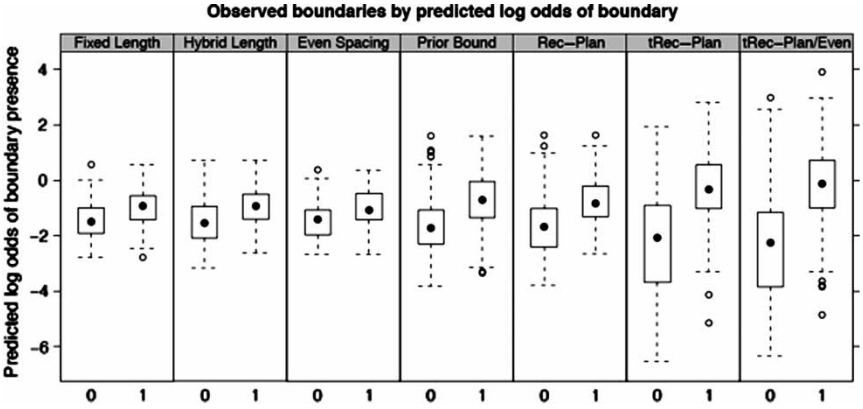

Figure 2. Box-and-whisker plot of the predicted boundaries for each model in Table 4 against ToBI boundary labels. The dot is the mean; boxes represent the inter-quartile range; dots represent values more than 1.5 times the distance of inter-quartile range from the mean. The *x*-axis indicates potential boundary locations in a trial (3–6 instances per trial) which were produced with or without a boundary, as determined by ToBI break indices. The *y*-axis indicates the log odds with which a given model predicts a boundary at one such potential boundary location. Better model fit is reflected by greater separation between boxes.

inter-quartile ranges have almost no overlap, suggesting that there is a strong relationship between the predictions of the thresholded recovery–planning model and the occurrence of boundaries, as coded by ToBI labels. Specifically, the phonological phrase boundaries which are assigned high log-odds of being a boundary based on the thresholded recovery–planning model were identified as boundaries by the ToBI labeler on a high proportion of trials. Similarly, the phonological phrase boundaries which were assigned low log-odds of being a boundary based on the thresholded recovery–planning model were much less likely to be identified as boundaries by the ToBI labeler.

Determining grain size

The analyses thus far demonstrate that the thresholded recovery–planning model is the best of the six models that were considered. Within this model, the size of material that has been produced and the material to be produced is measured in terms of phonological phrases. Because phonological phrases vary in length (in general and in our materials), using words or syllables as the size metric could result in an even better fit for the thresholded recovery–planning model. Consequently, we computed word/syllable weights at each phonological phrase boundary in (4) accordingly. For example, at the boundary between *the chapter* and *to the students of social science* in (4c), the left-hand weight, in phonological phrases, is one and the right-hand weight is two. The corresponding word weights are two and six, and the corresponding syllable weights are three and nine.

We then used word-based thresholded recovery–planning and syllable-based thresholded recovery–planning models to predict boundaries as described above. The results of these models are presented in Table 5. In modelling the acoustic data, the word-based thresholded recovery–planning model performs slightly better than the phonological phrase-based thresholded recovery–planning model ($z = 2.09$, $p < .05$); the syllable-based thresholded recovery–planning model also performs slightly better ($z = 2.19$, $p < .05$). In modelling the ToBI data, the word-based thresholded recovery–planning model outperforms the phonological phrase-based thresholded recovery–planning model ($z = 3.36$, $p < .0001$); the syllable-based thresholded recovery–planning model also outperforms the phonological phrase-based thresholded recovery–planning model ($z = 3.09$, $p < .0001$).

Attempting to integrate meaning constraints and balancing constraints

We have so far considered meaning-based pressures and balancing pressures as separate constraints on intonational boundary production. However, these constraints may apply together, perhaps independently, to determine

TABLE 5
Summary of the fixed effects in six mixed-effects models—three predicting intonational boundaries as determined by ToBI labels ($N = 2{,}961$), and three predicting phrase-final word duration and silence, ($N = 2{,}961$). The fixed effects in the models are the weights of the thresholded recovery-planning model of phrasing where size is determined by (a) the number of phonological phrases, (b) the number of words in the phonological phrases, or (c) the number of syllables in the phonological phrases. See the text for details

Models predicting ToBI boundaries						
Fixed effects	*Coefficient*	*SE*	*z*	*p*	*log-likelihood*	r^2
Intercept	−5.09	0.30	16.96	<.0001	−1179	.22
tRecovery-planning—phonphrase	1.43	0.08	18.73	<.0001		
Intercept	−5.00	0.29	17.21	<.0001	−1085	.29
tRecovery-planning—words	0.57	0.03	22.09	<.0001		
Intercept	−4.91	0.29	17.06	<.0001	−1097	.28
tRecovery-planning—syllables	0.36	0.02	21.75	<.0001		
Models predicting acoustic measures						
Intercept	0.41	0.02	23.57	<.0001	1435	.17
tRecovery-planning—phonphrase	0.05	0.002	24.55	<.0001		
Intercept	0.42	0.02	24.44	<.0001	1408	.21
tRecovery-planning —words	0.02	0.00	23.24	<.0001		
Intercept	0.41	0.17	24.07	<.0001	1442	.21
tRecovery-planning—syllables	0.01	0.00	24.95	<.0001		

intonational phrasing. A model which combines the predictions of the recovery–planning model with those of balancing models may therefore provide a better fit of the data. To investigate this possibility, we conducted two analyses, using both the thresholded recovery–planning model (the best model overall) and the even-spacing model (the best of the balancing models) to predict boundaries. First, the weights of the thresholded recovery–planning and even-spacing models were used to predict the duration of the phonological phrase-final word and the following silence in a linear mixed-effects model. Both sets of weights were included as fixed effects, with no interaction term, and subjects and items were included as random effects. The results of this model, presented in the bottom rows of Table 3, demonstrate that the inclusion of the even-spacing weights does not improve the model's predictive accuracy. Although the correlation between predicted and observed values of the thresholded recovery–planning balancing model is marginally higher than that of the thresholded recovery–planning model ($z = 1.87$, $p = .06$), the coefficient for the balancing component in this model is negative. The interpretation of such a model is therefore

one that over-weights the thresholded recovery–planning predictions, and subtracts a component corresponding to even spacing in order to predict the acoustic values. This is not the interpretation associated with the balancing hypotheses.

In a second analysis, the weights for the thresholded recovery–planning and even-spacing models were regressed against ToBI boundary presence in a linear mixed-effects logit model. As in the first analysis, both sets of weights were included as fixed effects, with no interaction term, with subjects and items as random effects. The results of this model, presented in the bottom rows of Table 4, demonstrate that the inclusion of the balancing weights does not improve the model's fit of the data over and above that of the thresholded recovery–planning alone. The correlation between predicted and observed values of the thresholded recovery–planning balancing model also did not differ significantly from that of the thresholded recovery–planning model ($z < 1$). In summary, we conclude that the balancing hypotheses that we evaluated here do not explain any of the prosodic segmentation effects in our data set.

GENERAL DISCUSSION

A data set consisting of sentences with three postverbal constituents of varying lengths was used to evaluate two classes of proposals for prosodic segmentation: (1) meaning-based proposals, such as Watson & Gibson's (2004) recovery–planning hypothesis; and (2) balancing proposals, whereby people produce boundaries at evenly spaced intervals. The locations of intonational boundaries were predicted best by a thresholded version of the recovery–planning hypothesis, which accounted for 22% of the variance in boundary placement when boundaries were defined acoustically (in terms of the duration of the phonological phrase-final word and any following silence), and 17% of the variance when boundaries were defined in terms of ToBI break indices. The fit of this model improved further when a more fine-grained measure of constituent length was used—words or syllables—rather than the coarse-grained measure of phonological phrases.

The best of the balancing models accounted for at most 9% of the variance, and did not improve the fit of the recovery–planning hypothesis when both sets of weights were included in a model evaluation. Thus, there does not appear to be any support for balancing models in the current data set. In the following section, we describe some ways in which the balancing algorithms failed to correctly predict speaker's boundary placement, and contrast these predictions with those of the recovery–planning model.

There are two places where fixed phrase length consistently over-predicts the presence of a boundary in our corpus (see Table 1). First, as discussed

above, it predicts a high probability of a boundary after the verb, because this location is two phonological phrases into the sentence. The results (see Table 2) demonstrate that boundaries are extremely rare in this location, occurring between only one and three per cent of the time (consistent with the predictions of the recovery–planning hypothesis, where boundaries are disallowed between verbs and their argument). Second, when the direct object is short and the indirect object is long (as in (4c) and (4g)), fixed phrase length predicts a high probability of a boundary following the first phonological phrase of the indirect object (e.g., after *The professor assigned the chapter to the students* and before *of social science*). The results demonstrate that boundaries occurred in this location only 43% (condition c) and 38% (condition g) of the time. This result is more consistent with recovery–planning, which predicts a low probability of a boundary in this location, because such a boundary would separate a head and its modifier (e.g., *the students, of social science*). Fixed phrase length also under-predicts boundaries in many locations. Because fixed phrase length predicts a boundary every two phonological phrases, it also predicts that boundaries should *not* occur after an odd number of phonological phrases. Consequently, this model predicts that a position after the fifth phonological phrase should be a poor place for a boundary. But many of these locations are locations where boundaries are likely to occur. In four conditions—(4b), (4c), (4f), and (4g)—this location separates the indirect object of the verb and a VP modifier, e.g., it separates *to the students of social science* and *yesterday* in (4c). When one of these two phrases is two or more phonological phrases in length—as in conditions (4c), (4f), and (4g)—there is a high likelihood of a boundary—.71, .70, and .87 of the time—in contrast to the prediction of the fixed phrase length proposal. Only when both the indirect object and the VP modifier are short (one phonological phrase), is there a low likelihood of a boundary in this position (.39 in condition (4b)). These data fit the predictions of the recovery–planning proposal, not the fixed phrase length proposal.

The hybrid fixed-length proposal differs from fixed phrase length in that, like recovery–planning, it correctly predicts that boundaries should not occur immediately following the verb in our materials. However, hybrid length over-predicts and under-predicts boundaries in much the same way that the fixed-length proposal does. Consider a position two phonological phrases following the verb. The hybrid length proposal predicts that this is never a good location for a prosodic boundary. But people do put a boundary in this location if it happens to be a constituent boundary, as predicted by the recovery–planning model. For example, in conditions (4a), (4b), (4d), (4e), (4f), and (4h), this location is a constituent boundary, and there is a correspondingly high likelihood for a boundary at these locations in our data: .59, .81, .84, .74, .78, and .81, respectively. Only when this location is

not a constituent boundary—as in (4c) and (4g)—is there a much lower likelihood of a boundary—.43 and .38, respectively. This is not the pattern predicted by the hybrid fixed-length hypothesis. Similarly, consider a position three phonological phrases following the verb. The hybrid fixed-length proposal predicts that this is always a good location for a prosodic boundary. However, people only tend to put a boundary in this location when it happens to be a constituent boundary, as predicted by the recovery–planning model. For example, in conditions (4d) and (4h), this location is not a constituent boundary, and there is a correspondingly low likelihood for a boundary at these locations: .45 and .41, respectively. In conditions (4b), (4c), (4f), and (4g) on the other hand, this location is a constituent boundary, and there is usually a correspondingly high likelihood for a boundary at these locations: .39, .71, .70, and .87. The only exception here is (4b), where both the constituent immediately preceding the boundary and the one immediately following are short: only one phonological phrase each. Thus the predictions of the recovery–planning model are a much better fit to the data than the hybrid length proposal.

Finally, the even-spacing proposal makes incorrect predictions in similar ways that the fixed phrase length and hybrid fixed-length proposals make incorrect predictions. Like the other proposals, the details of the predictions of the even-spacing proposal depend on the parameters of that proposal. Recall that the even-spacing model that works best was one that assumes (a) a threshold of two phonological phrases; (b) that leftover material gets produced last; and (c) that boundaries are disallowed between verbs and their obligatory arguments. This proposal works best in our corpus in part because (1) it disallows boundaries between verbs and their direct objects; and (2) it generally places a boundary before the sentence-final adverbial modifier, a location where a boundary often occurred. While this model made these correct predictions, it made many incorrect predictions. For example, this model predicts that the most likely locations for boundaries in condition (4b) (*The professor assigned the chapter on local history to the students yesterday*) will be following *chapter* and *students*. Contrary to this prediction, people are most likely to place a boundary between *the chapter on local history* and *to the students* (as predicted by the recovery–planning proposal). As a second example, the best even-spacing model predicts that the most likely locations for boundaries in condition (4d) (*The professor assigned the chapter on local history to the students of social science yesterday*) will be following *chapter* and *students* (the same predictions as for (4b)). Contrary to this prediction, people are most likely to place a boundary between *the chapter on local history* and *to the students of social science* and between *the students of social science* and *yesterday* (as predicted by the recovery–planning proposal).

It is worth noting that although the balancing models were not successful in explaining boundary placement in our materials, balancing may still play some role in prosodic segmentation. One possible explanation for the poor performance of balancing models in this data set is that balancing parameters may be different across individuals (e.g., a fast talker may balance with a threshold of four phonological phrases, whereas a slow talker may balance with a threshold of two phonological phrases). If this is the case, then using an "average" balancing parameter would not work well in accounting for patterns in boundary placement. This is in contrast to the meaning-based proposal, which uses the syntax/semantics of the utterance to predict boundaries, and which accurately predicts the average probability of producing a prosodic boundary across the different possible locations, as can be seen from Table 2. Unlike balancing information, meaning-based grouping—derived from the syntax/semantics of the language—is relatively invariant across native speakers of the language. Even so, the best meaning-based proposal only accounted for 22% of the variance in duration cues associated with boundary production. Consequently, there is still a lot of variance left unaccounted for in this model. The remaining unexplained variance is plausibly due to individual differences in participants' boundary production. Future work could investigate whether more variance might be explained by including individual-specific balancing parameters in modelling the data across subjects.

In conclusion, W&G's (2004) meaning-based proposal was shown to account for a substantial amount of variance in boundary productions. The success of the model is plausibly due to meaning-based grouping being relatively invariant across native speakers. In contrast, we have found little support for balancing hypotheses. The failure of these models may be due to the possibility that individuals differ in their balancing parameters, so that average balancing parameters do not improve model performance overall.

Manuscript received May 2009
Revised manuscript received July 2010
First published online 20 September 2010

REFERENCES

Albritton, D., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 714–735.

Allum, P., & Wheeldon, L. R. (2007). Planning scope in spoken sentence production: The role of grammatical units. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*, 791–810.

Arnold, J., & Tanenhaus, M. (2010). Disfluency effects in comprehension: How new information can become accessible. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

Augurzky, P. (2008). Prosodic balance constrains argument structure interpretation in German. Poster presented at the 14th AMLaP conference, Cambridge, UK.

Bachenko, J., & Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, *16*(3), 155–170.

Beckman, M., & Ayers Elam, G. (1997). Guidelines for ToBI labeling (Version 3). Manuscript and accompanying speech materials, Ohio State University. [Obtain by writing to tobi@ling.ohio-state.edu.]

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*, 145–204.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer (Version 4.3.10) [Computer program]. Retrieved from http://www.praat.org/

Breen, M., Dilley, L., Kraemer, J., & Gibson, E. (2010). *Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)*. Manuscript submitted for publication.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, *109*, 274–280.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*, 201–242.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Dehé, N. (2002). *Particle verbs in English: Syntax, information structure, and intonation*. Amsterdam and Philadelphia, PA: John Benjamins.

Dilley, L., Breen, M., Gibson, E., Bolivar, M., & Kraemer, J. (2006). A comparison of inter-coder reliability for two systems of prosodic transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). Proceedings of the international conference on spoken language processing, Pittsburgh, PA.

Ferreira, F. (1988). *Planning and timing in sentence production: The syntax-to-phonology conversion*. Unpublished dissertation, University of Massachusetts, Amherst, MA.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*, 210–233.

Ferreira, F. (1993). The creation of prosody during sentence production. *Psychological Review*, *100*, 233–253.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*, 1151–1177.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57–84.

Fodor, J. (1998). Learning to parse? *Journal of Psycholinguistic Research*, *27*, 285–319.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production. Vol. 1: Speech and talk* (pp. 177–220). London: Academic Press.

Gee, J. P., & Grosjean, F. (1983). Performance structures. A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures insentence production. *Cognitive Psychology*, *11*, 58–81.

Hirotani, M. (2007). Prosody and LF interpretation: Processing Japanese Wh-questions. *Phonological Studies*, *10*, 67–68.

Hirschberg, J. (1995, August). Prosodic and other acoustic cues to speaking style in spontaneous and read speech. In *Proceedings of ICPhS* (Vol. 2, pp. 36–43). Stockholm.

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.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140.

Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50*, 194–231.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, *7*, 102–122.

Lehiste, I., Olive, J., & Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of Acoustical Society of America*, *60*(5), 1199–1202.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris Publications.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Unpublished dissertation, MIT.

Pitrelli, J., Beckman, M., & Hirschberg, J. (1994). Evaluation of prosodic transcription labelling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 1, pp. 123–126).

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America*, *90*, 2956–2970.

Sanders, E., & Taylor, P. A. (1995). Using statistical models to predict phrase boundaries for speech synthesis. In *Proceedings of Eurospeech '95* (pp. 1811–1814). Madrid.

Schafer, A., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169–182.

Selkirk, E. O. (1984). *The relation between sound and structure.* Cambridge, MA: MIT Press.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., . . . Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the international conference on spoken language processing* (pp. 867–870). Banff, Canada.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential contest. *Journal of Memory and Language*, *48*, 103–130.

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

Streeter, L. (1978). Acoustic determinants of phrase boundary perception. *Journal of Acoustical Society of America*, *64*, 1582–1592.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 1–28). Berlin and New York: De Gruyter.

Wang, M. Q., & Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, *6*, 175–196.

Watson, D., Breen, M., & Gibson, E. (2006). The role of syntactic obligatoriness in the production of intonational boundaries. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(5), 1045–1056.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*, 713–755.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of Acoustical Society of America*, *91*(3), 1707–1717.

APPENDIX 1

1. The mobster paid the bounty (of thirty diamonds) to the gangster (with burly henchmen) quickly/before the crime was committed.
2. The caterer brought the pastries (with lemon filling) to the party (for Oscar winners) early/before the guests had arrived.
3. The gigolo sent a bouquet (of sixty roses) to the showgirl (from Hello Dolly) on Sunday/before the performance last night.
4. The colonel assigned the mission (of killing Castro) to the soldier (with sniper training) last night/last night at the Pentagon.
5. The wizard granted the powers (of magic healing) to the suitor (of England's princess) last night/after being threatened with death.
6. The matriarch left the necklace (with sapphire inlay) to the daughter (of peasant parents) secretly/before the family found out.
7. The director offered the payment (of thirty million) to the actor (of poignant dramas) yestreday/after filming had already begun.
8. The academy presented the award (of greatest import) to the actor (of little renown) on Sunday/last week in Los Angeles.
9. The executive delivered the statement (of corrupt actions) to the judges (of business conduct) regretfully/before a ruling was issued.
10. The professor assigned the chapter (on local history) to the students (of social science) yesterday/after the first midterm exam.
11. The writer pitched the story (of happy orphans) to the chairman (of Disney Studios) at lunch/over several drinks after lunch.
12. The student gave the basket (of chocolate brownies) to the teacher (of ancient history) today/before the start of vacation.
13. The lieutenant evacuated the soldiers (of several platoons) to a region (with unarmed locals) yesterday/after the mysterious phone call.
14. The girl attached the posters (of missing children) to the windows (of local buildings) today/after her shopping trip downtown.
15. The priest delivered the turkeys (with homemade stuffing) to the homeless (at local shelters) on Thursday/before people arrived for dinner.
16. The socialite donated the suitcase (of lovely dresses) to the woman (in dirty clothing) yesterday/after meeting her at church.
17. The lawyer left the duties (of mindless errands) to the partner (with lower status) this morning/after the lengthy conference call.
18. The girl lent the booklet (of practice exams) to the classmate (from second period) on Friday/before the test on Friday.
19. The gentleman sent the bouquet (of gorgeous roses) to the woman (with shiny lipstick) on Monday/after spotting her from afar.
20. The millionaire assigned a chauffeur (with little patience) to his mistress (in Southern Europe) today/after a quarrel on Friday.
21. The station offered the ballad (with minor changes) to the public (in nearby cities) last week/after the debate last week.
22. The grandmother gave the necklace (of twenty pearls) to the grandson (from Kansas City) on Sunday/at the annual family reunion.
23. The architect placed the statue (of Roger Sherman) in the courtyard (with pretty flowers) carefully/with tremendous pride and satisfaction.
24. The son put his backpack (with heavy textbooks) in the kitchen (with seven people) last night/without stopping to eat dinner.
25. The critic handed the letter (for Steven Spielberg) to the postman (with curly sideburns)

personally/in the sunshine of morning.

26. The committee allocated the money (from Tuesday's auction) to the members (from Costa Rica) yesterday/after numerous hours of discussion.
27. The bride put the favours (of mini bouquets) on the tables (of several guests) happily/before the wedding reception began.
28. The spy told the secrets (of deadly weapons) to the leaders (of foreign nations) quietly/through a network of operatives.
29. The salesman conveyed his advice (on buying vases) to the clients (from rural Texas) on Friday/after a meeting on Friday.
30. The professor assigned a project (on Asian Studies) to his students (with heavy workloads) yesterday/without regard for other classes.
31. The tycoon lent the limo (with leather seating) to his buddies (from Swarthmore College) often/for several days last month.
32. The referee explained the format (of soccer contests) to the players (from Amherst College) on Friday/before the big tournament began.