Squeezing through the Now-or-Never bottleneck: Reconnecting language processing, acquisition, change and structure

Nick Chater & Morten H. Christiansen

Abstract

If human language must be squeezed through a narrow cognitive bottleneck, what are the implications for language processing, acquisition, change, and structure? In our target article, we suggested that the implications are far-reaching and provide an integrated account of many apparently unconnected aspects of language and language processing, as well as suggesting revision of many existing theoretical accounts. With some exceptions, commentators were generally supportive both of the existence of the bottleneck and its potential implications. Many commentators suggested additional theoretical and linguistic nuances and extensions, links with prior work, and relevant computational and neuroscientific considerations; some argued for related but distinct viewpoints; a few, though, felt traditional perspectives were being abandoned too readily. Our response attempts to build on the many suggestions raised by the commentators and to engage constructively with challenges to our approach.
R1. Introduction

In our target article, we argued that a powerful and general cognitive constraint, the Now-or-Never bottleneck, has far-reaching consequences for both language comprehension and production. This perspective implies that language acquisition and language change proceed construction-by-construction, rather than involving more abrupt, system-wide shifts. We argued, moreover, that the picture that arises from the Now-or-Never bottleneck has implications for the structure of language itself: syntactic structure is viewed as processing history, thus enforcing a tight link between the psychology of language processing and linguistic theory.

The Now-or-Never bottleneck is a general cognitive constraint that, we suggest, applies to perception, motor control, reasoning and memory: unless information is recoded and/or used rapidly, it is subject to severe interference from an onslaught of further information. Our article explores possible implications of the Now-or-Never bottleneck for language: how it is processed and acquired, how languages changes, and the structure of language of itself. The argument is that the Now-or-Never bottleneck has profound implications in each of these domains: for example, it requires that processing is incremental and predictive, using a Chunk-and-Pass mechanism; that acquisition is item-based; that languages changes construction-by-construction; and that there may be an intimate relationship between language structure and processing.

The commentators on our article have provided a rich variety of perspectives and challenges with which to evaluate and potentially to further develop this account. We have grouped our response to commentators according to key themes that emerge.

The first theme, The status of the Now-or-Never bottleneck concerns the evidence for, and nature of, the bottleneck. Key questions include: Does the psychological and linguistic evidence support (Ferreira & Christianson; Kempson, Chatzikyriakidis & Cann; Potter) or contradict (Baggio & Vicario; Chacón, Momma & Phillips; Endress & Katzir) the existence of the bottleneck? Have we overstated its scope (Levinson, MacDonald)? What is its neural basis (Frank & Fitz; Grossberg; Honey, Chen, Musch & Hasson; Huyk)? How can the hypothesis be elaborated (Dumitru; Potter)? And if we accept the existence of the Now-or-Never bottleneck, should it be treated as basic, or as arising from more fundamental principles (e.g., Badets; Bicknell, Jaeger & Tanenhaus; Lotem, Kolodny, Halpern, Onnis & Edelman; Wilkinson)?

A second set of issues, discussed in The case for Chunk-and-Pass Language Processing, focuses on the empirical and computational viability of the framework for language processing that we derive from the Now-or-Never bottleneck. According to the Chunk-and-Pass framework, language comprehension requires a succession of increasingly abstract chunking operations and, at each level, chunking must occur as rapidly as possible, and the resulting chunks immediately passed to higher levels. The reverse process, where the speaker converts an abstract message into articulatory instructions, is proposed to involve what we term Just-in-Time language production.
Key questions include: how does the Chunk-and-Pass framework relate to existing theories of language processing, both in psycholinguistics (Bicknell et al.; Chacón et al.; Ferreira & Christianson; MacDonald; O’Grady) and computational linguistics (Huyck and Gómez-Rodríguez)? How do these proposals relate to experimental data (Baggio & Vicario; Healey, Howes, Hough & Purver), including effects of top-down processing (Dimutru; Healey et al; MacDonald; Potter)? Can our account meet the challenges of interactive dialogue (Badets; Baggio & Vicario; Healey et al.; Kempson et al.; Levinson)? How far does the Chunk-and-Pass approach apply to sign language (Emmorey), and to non-linguistic domains such as music and action (Lakshmanan & Graham; Maier & Baldwin)?

A third set of issues, *Consequences for language acquisition, evolution and structure*, concerns the implications of the Now-or-Never bottleneck and Chunk-and-Pass processing for language acquisition, evolution, and structure. In our target article, we argued that the bottleneck has far-reaching implications for language across multiple timescales, ranging from the duality of patterning observed across languages (roughly, having distinct phonological and lexical levels), the locality of most linguistic regularities, and what we take to be the instance-based nature of language acquisition and language change. Key questions include whether our account provides sufficient constraints to explain language acquisition (Endress & Katzir; Lakshmanan & Graham; Wang & Mintz) and how it may be developed further (Lewis & Frank; Maier & Baldwin); and how far can the account explain language change and evolution (Behme; Bergmann, Dale & Lupyan; Endress & Katzir; Lewis & Frank; Lotem et al.)? Some commentators explore how this approach can be a productive framework for understanding regularities within and across languages (Kempson et al; O’Grady); while other believe that further constraints are required (Bever, Piatelli-Palmerini & Medeiros; Chacón et al.; Endress & Katzir; Wang & Mintz).

In the remainder of this response to commentators, we will discuss these three sets of issues in turn, before drawing general conclusions and considering directions for future work.

**R2. The nature of the Now-or-Never bottleneck**

Memory is fleeting: sensory and linguistic information is subject to severe interference from the continual onslaught of new material. If the input is not used or recoded right away, it will be lost forever: this is the Now-or-Never bottleneck.

For this reason, our memory for arbitrary sequences of sounds is extraordinarily limited (Warren, Obusek, Farmer, & Warren, 1969). Yet we are able to process highly complex non-arbitrary sequences of linguistic input (and, similarly, musical input and action sequences). We proposed that these observations imply that sensory and linguistic information must be used or recoded into higher-level representations right away, to avoid being lost forever.

What is the origin of the Now-or-Never bottleneck? In our target article, we stressed the importance of interference—new input interferes with existing input, particularly between
elements that overlap phonologically or semantically. Such interference has been observed at a wide variety of representational levels in studies of memory for serial order (Brown, Neath & Chater, 2007). Likewise, as noted by MacDonald, sentences containing words with overlapping phonological forms and meaning create processing problems (e.g., ‘The baker that the bakerer sought bought the house’ vs. ‘The runner that the banker feared bought the house’, Acheson & MacDonald, 2011; see also, Van Dyke & Johns, 2012, for a review).

Another possible origin of the bottleneck stems not from interference, but from one or more capacity limited buffers (discussed by Levinson). So, for example, Miller (1956) famously suggested a capacity of 7±2 chunks in short term memory, an approach enriched and updated by Baddeley and colleagues (e.g., Baddeley, 1992; Baddeley & Hitch, 1974). More recently, Cowan (2000) argues for a capacity limit of 4±1 items. Our reading of the recent memory literature is that many, and perhaps all, aspects of memory limitations may best be understood in terms of interference rather than limited capacity buffers, because the same patterns of forgetting and memory errors are observed over many timescales (e.g., Brown, Neath & Chater, 2007). From this perspective, apparent capacity limitations are a side-effect of interference, rather than stemming from, for example, a fixed number of ‘slots’ in memory (see also Van Dyke & Johns, 2012).

From the point of view of the target article, the key issue is the limited nature of the bottleneck, whether it stems primarily from interference, capacity limitations or a combination of the two. Note, in particular, that memory performance depends on the number of chunks involved, and what counts as a chunk depends on prior experience with relevant material. Hence the same sequence of phonemes may, over experience, be chunked into a series of syllables or words, or into a single multiword chunk (Jones, 2012). We stress, too, that interference effects will operate between chunks—i.e., chunks are not merely encapsulated units—so that some of the internal structure of chunks will be retained. This is evident, for example, in phonological interference effects in memory for serial order (Burgess & Hitch, 1999). Thus, although some commentators (e.g., Bicknell et al., MacDonald) seem to have taken our notion of “lossy compression” as indicating a near total loss of information, we use the term in the standard computer science sense as indicating that not all information is retained. More generally, we are able to outline the consequences of the Now-or-Never bottleneck without taking a stand on the exact nature of the underlying memory representations—although, of course, within the general framework developed here, more detailed memory models will allow for more fine-grained predictions about language processing.

Indeed, we suggest that one fruitful direction for research is to explore cognitive models in which processing and memory are not distinct mechanisms. As Honey et al. point out, it may be appropriate to see memory as arising from on-going neural processing activity, rather than as located in distinct stores (see, e.g., Crowder, 1993; Kolers & Roediger, 1984). From this viewpoint, processing and memory operations should be located in the same brain regions (Hasson, Chen & Honey, 2015). This perspective has also been applied to accounts of individual
differences in language processing, modeled using Simple Recurrent Networks (Elman, 1990), in which the same connections and weights store and process linguistic input (MacDonald & Christiansen, 2002). This type of model captures the relationship between language processing and short-term memory performance, without any functionally distinct working memory (by contrast with, for example, production system models such as Just and Carpenter’s [1992] CC-READER). As we shall discuss further below, in this integrated perspective on memory and processing it is not possible to modify memory capacity independently of processing operations (Christiansen & Chater, 2015, 2016; MacDonald & Christiansen, 2002). Thus, memory capacity is not a free parameter that can be independently selected for by natural selection (see our discussion of Lotem et al.).

Honey et al. underscore our claim that the Now-or-Never bottleneck implies longer integration timescales for more abstract levels of representation. They substantiate this view with evidence from fMRI and intracranial recordings, countering Vicario & Baggio’s concern that our multilevel representational approach lacks neural foundations. According to Honey et al., incoming information is continually integrated with prior information—yet once integration has occurred, the resulting interpretation and knowledge updating becomes entrenched and difficult to revise (Ferreira & Christianson). Consistent with such interpretative entrenchment, Tylén, Christensen, Roepstorff, Lund, Østergaard and Donald (in press) found that when a narrative had a coherent storyline, then incidental facts tended to be forgotten if they were not central to the plot. However, when the storyline was jumbled, there was a greater recall of incidental semantic facts, presumably because integration was not possible. Importantly, an fMRI version of the same experiment yielded activation of the same cortical hierarchies, from lower-level sensory circuits to higher-level cognitive areas, as noted by Honey et al. (and discussed in the target article).

R.2.1 Challenges to the Now-or-Never bottleneck

Several commentators question the severity of the Now-or-Never bottleneck. Some of these concerns, however, focus on consequences that do not follow from the bottleneck. For example, as illustrated by SF’s spectacular memory for sequences of numbers chunked by running times, chunking low-level material facilitates memory for that material. More broadly, low-level information is only remembered to the extent that it has been processed. So the Now-or-Never bottleneck does not imply complete amnesia for past low-level sensory or linguistic information—people can, after all, remember tunes and poems by heart. What they cannot do is recall unprocessed sequences of noises or letters, which they are unable to chunk in light of prior experience. So while we can remember new words in our language, recalling a complex sound-pattern from a foreign language (e.g., for speakers of English, a word or phrase in Khoisan) will be very difficult. Hence, Endress & Katzir’s claim that children can learn a word from a single encounter does not challenge the Now-or-Never bottleneck (the notion of fast-mapping has,
though, been questioned in some recent studies, e.g., Horst & Samuelson, 2008; McMurray, Horst & Samuelson, 2012).

We stress also (pace Endress & Katzir) that the bottleneck applies equally to explicit and so-called implicit memory (i.e., with or without awareness), if indeed such a distinction can be defended (e.g., Shanks & St. John, 1994). Our claim is that memory is dependent on processing, and this remains true irrespective of whether memory is assessed through explicit or implicit measures. For example, many psychology undergraduates will have been exposed to the hard-to-see ‘Dalmatian’ (see, e.g., Gregory, 2005). Famously, once one can see the pattern as a Dalmatian, the Dalmatian interpretation is typically available many years later (e.g., to help segment the image, an implicit measure of memory)—and the image will immediately be recognized as familiar and as a Dalmatian (explicit measures). But, of course, people who have not successfully found the Dalmatian ‘Gestalt’ will, of course, not remember that they have seen this specific pattern of black and white marks on a piece of paper or a computer screen many years before. In short, an image is only memorable to the extent that it has been successfully processed. This explains why prior exposure to an image will assist the processing of later copies of the same image, because such exposure helps create a ‘gist’ that can be re-used, allowing for cumulative learning effects over multiple exposures (see, for example, Endress & Potter, 2014).

Similarly, Bicknell et al. stress that perceptual data is not necessarily immediately forgotten—and we agree. The Now-or-Never bottleneck implies that perceptual or linguistic data that cannot be successfully processed into higher-level representations will suffer severe interference from subsequent material. But where that data can be recoded successfully, more low-level details may be retained, because they are embedded within a richer memory structure, thus countering interference from subsequent material to some extent. Nonetheless, we would anticipate that recalling such low-level details is likely to be cognitively effortful, although some details may be retained when crucial to the task in hand.

This point is illustrated by a study that Bicknell et al. describe, by Connine, Blasko and Hall (1991), employing a phoneme labeling task. Participants indicate which of two sounds they heard at the beginning of the third word in a sentence, and are instructed to use any available information from the sentence to make their response. The stimuli were ambiguous between a voiced and unvoiced initial consonant, yielding a blend of dent and tent, followed by a disambiguating context: “When the ___ in the fender/forest...” Therefore, while encoding the word, participants are explicitly instructed to pay attention to the details of the first phoneme. So some low-level information is likely to be retained over a short period. Bicknell et al. report their own study indicating slightly longer periods of retention of phonemic information, over six syllables, when participants are refrained from responding till the end of the sentence. But this hardly changes the broad message that the “raw” sensory input is rapidly lost, presumably through interference, although some limited information can, as we would predict, be retained through being encoded in larger units (e.g., through retaining a memory of the degree of ‘ambiguousness’ of the word dent or tent).
Note that Connine et al. highlight task-specific effects as a possible driver of their results: “One major issue left unresolved by the present research is the degree to which delayed commitment is subject to strategic factors introduced by task specific demands.” (p. 246). With this in mind, we can only agree, with Bicknell et al. (and also Ferreira & Christianson), that memory (including memory for low-level information encoded into higher level units) can be used strategically in the service of task goals (e.g., Anderson & Milson, 1989; Anderson & Schooler, 1991). Indeed, as noted by Potter, our framework seems naturally compatible with allowing newly built structures “to be influenced by the observer’s task or goal” (Potter, p.000). Moreover, it is possible that such strategic task-related effects may appropriately be modeled by bounded rational analysis, as Bicknell et al. suggest. Similarly, we suggest that this approach to modeling task-specific effects is compatible with the ‘good enough’ processing model described by Ferreira & Christianson (Ferreira & Swets, 2002). We see the Now-or-Never viewpoint as providing a framework within which ‘boundedly rational’ and ‘good enough’ models may fruitfully be integrated.

Whereas Endress & Katzir and Bicknell et al. stress, and we agree, that not all low-level information is lost immediately (though it will be lost if it cannot be processed into higher level units), Baggio & Vicario argue that the processing of sequential material such as language should not be viewed as a race against time at all. They do not deny the existence of the Now-or-Never bottleneck, but suggest that the brain has a number of mechanisms through which the effects of the bottleneck can be countered, including inference, pragmatics, and skills associated with literacy.

Yet we are not sure that Baggio & Vicario’s suggestions change the picture substantially. Focusing for now on reading, even though we can always re-fixate a word that we have missed or misread while reading, becoming a fluent reader requires overcoming a reading-based analogue of the Now-or-Never bottleneck: 1) memory for visual information is short-lived (60-70 msec, Pashler, 1998); 2) visual input is taken in at a fast rate during normal reading (about 200 words per minute, Legge et al., 1985); and 3) memory for visual sequences is limited (to about 4 items; Luck & Vogel, 1997). Because memory for what has just been read is short-lived and subject to rapid interference, we suggest that readers must perform chunking operations on text input as quickly as possible in order to read fluently. Indeed, individual differences in chunking ability predict self-paced reading performance (McCauley & Christiansen, 2015b).

**R2.2 Is the Now-or-Never bottleneck a side-effect of a deeper constraint?**

In our target article, we argued that the Now-or-Never bottleneck provides a powerful motivation for on-line prediction in language processing, and cognition more broadly. Given the underspecified nature of the sensory and linguistic input, predictive information is required to analyze new input as rapidly as possible, before it is obliterated by the onslaught of further
material. Similarly, prediction is required for on-line learning, in which the disparity between predictions and sensory data can immediately be used to drive learning. According to the Now-or-Never bottleneck, unless the disparity between predictions and input is computed and exploited right away, the sensory information will be lost, and with it, the opportunity for learning.

By contrast, Wilkinson and Badets argue, from different perspectives, that on-line prediction should not be seen as helping to deal with the Now-or-Never bottleneck, but as the central engine of cognition. There might not be substantial disagreement here, however. A cognitive theory based on prediction still has to specify at which point the error between prediction and sensory or linguistic input is assessed, to guide action and shape learning. The Now-or-Never bottleneck requires that prediction error is calculated and used to drive learning on-line: if the disparity between prediction and sensory input is not calculated right away, then sensory input will be lost. Notice that, by contrast, many prediction-based learning methods do not learn on-line. For example, the parameters in connectionist networks or Bayesian models are often adapted to provide the best fit to the whole ‘batch’ of available data, which typically involves storing and resampling this data throughout learning. Indeed, the requirement for learning to be on-line is very strong: on-line learning algorithms face the danger of so-called ‘catastrophic interference’ where learning new items damages memories of old items (e.g., French, 1999). Such catastrophic interference can, as we note, be avoided by using item-based learning models, so that learning from experience involves not re-fitting the parameters of a model (e.g., a stochastic phrase structure grammar, or the like), but continually adding to, and then generalizing from, a data-base of stored exemplars (e.g., an inventory of constructions). Needless to say, sensory experience must be encoded in an abstract form (rather than purely as “raw” acoustic or visual input) to reduce interference with other stored items. In our target article, we argue that item-based learning is a plausible model for language acquisition (Tomasello, 2003); and the need for on-line predictive learning, imposed by the Now-or-Never bottleneck, may favor item-based learning throughout perception and cognition more broadly (e.g., Kolodner, 1993; Poggio & Edelman, 1990).

From a different theoretical viewpoint, Lotem et al. raise the possibility that the Now-or-Never bottleneck should not necessarily be viewed as a fixed constraint on cognitive machinery, but may instead itself be an adaptation of our learning mechanisms, driven by natural selection (see also Endress & Katzir’s discussion of Major & Tank, 2004). The argument of our target article focuses on the nature and implications of the Now-or-Never bottleneck; but the question of the origins of the bottleneck is, of course, of great interest. Lotem et al. argue that the bottleneck has been adapted through natural selection to optimize the brain’s ability to learn. They note that a wide variety of evidence shows that memory performance varies between individuals, and is to some extent heritable. They interpret this variation to suggest that the size of the bottleneck is itself variable—and that this size can potentially be selected for. This viewpoint would, for example, be compatible with theories of memory, mentioned earlier, in which memory consists
of one or more capacity-limited buffers (e.g., Baddeley, 1992)—and hence where the capacity limit can be adjusted (as is appropriate, for example, in thinking about computer RAM memory or hard disk capacity).

We suggest, by contrast, that human memory and processing are fundamentally integrated and that the Now-or-Never bottleneck arises from interference effects which are unavoidable, given that the same neural and computational machinery is used for successive, and potentially strongly overlapping, and hence interfering, inputs (e.g., Hintzman, 1988; Murdock, 1983; Brown, Neath & Chater, 2007). From this standpoint, the Now-or-Never bottleneck is not usefully characterized as having a variable size, which is subject to independent variation and selection. Rather, the bottleneck emerges from the computational architecture of the brain; and variation in memory performance depends on the effectiveness of Chunk-and-Pass mechanisms to mitigate its impact. So S.F.’s ability to encode streams of digits as running times, indicates not a particularly wide ‘bottleneck,’ but rather a particularly efficient recoding strategy (Ericsson, Chase, & Faloon, 1980). Expert chess players are able to recall positions of real chess games by encoding them using a rich set of “chunks” from prior games (yet even top chess players have no memory advantage for ‘nonsense’ chess positions and neither do they have significantly above average general visuo-spatial abilities, Simon & Chase, 1973; Waters, Gobet & Leyden, 2002). Similarly, we suggest that individual differences in the efficacy of language processing operations will depend on being able to draw on a rich set of prior linguistic experiences to efficiently recode linguistic input (Christiansen & Chater, 2016; Jones, 2012; MacDonald & Christiansen, 2002).

From this standpoint, it is not appropriate to see the size of the Now-or-Never bottleneck as a free parameter that can be optimized through selection and variation, as embodied in Lotem et al.’s variable “time-window” in their computer simulations (e.g., Kolodny, Edelman & Lotem, 2014, 2015a, 2015b). Note, too, that the “window” in this model is large (e.g., 50-300 items) compared to buffers typically postulated in the study of human memory (Baddeley, 1992), so its psychological status is not clear either.

In any case, to the extent that Lotem et al. see the Now-or-Never bottleneck for language as shaped specifically to the linguistic environment, their approach appears to depend on the structure of language being exogenously fixed, to provide a stable target for adaption of the Now-or-Never bottleneck. But language is not given exogenously, but is shaped by generations of rapid cultural evolution to fit with, among other things, the learning and processing biases of the brain, including the Now-or-Never bottleneck. We have suggested elsewhere that language is shaped by the brain, rather than the brain being shaped by language (Christiansen & Chater, 2008). So linguistic regularities will arise from, among other things, the Now-or-Never bottleneck; and hence the Now-or-Never bottleneck should be seen as prior to, rather than as an adaptation for, the structure of language.

**R2.4 Neural plausibility?**
How might the Now-or-Never bottleneck be implemented neurally? Grossberg argues that many key aspects of our approach are already embodied in existing computational models of neural function created by his research team, and, in particular, in the notions of Item-Order-Rank (IOR) working memory and by a learning and chunking mechanism called the Masking Field (MF) (for a less detailed discussion along somewhat similar lines see Huyck). We are sympathetic with the proposal that Chunk-and-Pass processing, and, more broadly, the serial character of high-level thought (e.g., Pashler, 1998), derives from the basic operating principles of the brain, as carrying out a sequence of parallel constraint satisfaction processes. The data outlined by Honey et al. suggest that each computational step (e.g., chunking and recoding linguistic input) may work in parallel across large areas of the brain, so that multiple processes at the same representational level cannot be carried out simultaneously, and hence language processing, and high-level thought more generally, is sequential (e.g., Rumelhart, Smolensky, McClelland & Hinton, 1986). If this is right, then the Now-or-Never bottleneck may be a side-effect of the basic principles of neural computation, rather than a free parameter than can be readily modified by natural selection (contra Lotem et al).

Frank & Fitz offer a very different perspective on brain function inspired by the processing properties of the cerebellum (Fitz, 2011). They question the severity of the bottleneck in light of computational results from what they term ‘reservoir computing,’ in which an untrained neural network projects a temporal input stream into a high dimensional space; a second network is trained to read off information from the ‘reservoir.’ They report simulations that they take to show that the network can reliably recover complex sequential input after long delays. Interesting as these results are, they seem to provide a poor fit with the large literatures on both human memory limitations and restrictions on language processing. It is thus unclear whether such networks would predict the aspects of language processing discussed in our target article, and by other commentators (e.g., Ferreira & Christianson; Grossberg; Kempson et al.).

R3 Chunk-and-Pass Language Processing

The Now-or-Never bottleneck is a fundamental constraint on memory that the language system deals with by Chunk-and-Pass comprehension and Just-in-Time production. The very phrase “Chunk-and-Pass” has, to some commentators, suggested a link with the Sausage Machine parsing model of Frazier and Fodor (1978). This has led some commentators to level concerns at the Chunk-and-Pass approach that are more appropriately directed at the Sausage Machine (Bicknell et al; Chacón et al.; Ferreira & Christianson; Healy et al.; MacDonald). According to the Sausage Machine model, a preliminary syntactic analysis is created within a window of about six words and then shunted off as a packet (like successive sausages coming out of a real sausage machine) to a second stage that completes the syntactic parsing. But while the Sausage Machine has a packet-by-packet character, it differs fundamentally from the Chunk-and-Pass model along at least three key dimensions. First, the Chunk-and-Pass account operates at a
variety of representational levels, using units that have been acquired by item-based learning—so Chunk-and-Pass processing is not restricted to the syntactic units used in parsing. Second, while the operation of the Sausage Machine is informationally encapsulated from semantic and pragmatic factors, the Chunk-and-Pass model assumes that all sources of information, from low-level sensory input to pragmatics and world knowledge are brought to bear on-line to create and recode chunks at all levels of analysis. Thus, we stress that the Chunk-and-Pass view includes top-down influences (see Dumitru), rather than operating purely bottom-up in a modular fashion (a concern raised by Healey et al., Lotem et al. and MacDonald). Third, note that, unlike the Sausage Machine, which postulates cognitively decisive break-points at the boundaries between ‘sausages’ (i.e., phrase structure created by the parser), the Chunk-and-Pass viewpoint allows links (and interference) between items which are not grouped within the same chunk (e.g., words which are not in the same phrase or clause). But the strength of such links will reduce rapidly, in a graded fashion, as the ‘distance’ between items increases, as would be predicted by the memory interference processes that we take to underlie the Now-or-Never bottleneck. Chunk-and-Pass processing implies a strong bias towards local structure in language, but is entirely compatible with the existence of some non-local dependencies (see Bever et al.; Healey et al.; Levinson). We emphasize that the Now-or-Never bottleneck explains the remarkably, though not completely, local structure of language (as noted by Kempson et al.), with its hierarchy of levels of representations largely corresponding to local sequences of linguistic material. As we outlined in our target article, this contrasts with the with batch-coded communication signals used in engineering and computer science, which are optimal from the point of view of information theory (Cover & Thomas, 2006).

Turning to production, we argued that the Now-or-Never bottleneck implies that once detailed low-level production instructions have been assembled, they must be executed right away, or they will be obliterated by interference from the on-coming stream of later instructions: this is Just-in-Time production. Some commentators (Chacón et al.; Ferreira & Christianson; MacDonald) have taken Just-in-Time production to imply so-called ‘radical incrementality,’ in which phonological words are articulated immediately in the absence of any planning ahead. They rightly noted that such radical incrementality is inconsistent with evidence of task-related effects on production. For example, Ferreira and Swets (2002) showed that participants plan ahead when producing utterances involving the results of arithmetic calculations. Indeed, speakers appear to plan beyond the immediate phonological word, but likely no more than a clause in advance (e.g., Bock & Cutting, 1992). We want to stress, though, that just as comprehension at the discourse level takes place over a relatively long timescale, so does planning at the discourse or conceptual level in production. This is because chunks at the discourse level have a longer duration than articulatory chunks (see Honey et al.). Whereas planning at the level of the phonological word may be quite short in temporal scope, planning will extend further ahead at the level of multiword combinations (what might traditionally be called the “grammatical level”), and even longer at the conceptual/discourse level (e.g., Smith & Wheeldon, 2004). Thus, the evidence that Chacón et al. discuss in this regard (e.g., Smith &
Wheeldon, 1999; Lee, Brown-Schmidt & Watson, 2013) is not inconsistent with Just-in-Time production.

Nonetheless, it is important to note that that people do interleave planning and articulation processes when producing utterances under time pressure (Ferreira & Swets, 2002). Given the speed of turn-taking (e.g., as noted by Levinson), such time pressures may be the norm in normal conversations, limiting the amount of advance planning possible. This is reflected by the patterns of disfluencies observed in production, indicative of brief planning ahead at the clausal level (e.g., Ferreira, 1993; Holmes, 1988). We see this limited planning ahead as compatible with Just-in-Time production, whereby production is limited to just a few chunks ahead for a given level of representation. Crucially, as noted in the target article, such chunks may involve multiword sequences, which are articulated as units rather than as a chain of individual words (Arnon & Cohen Priva, 2013; Bybee & Schiebman, 1999). This allows speakers to plan ahead to some degree when this is required by task demands, though our account suggests that such planning would be limited to a few chunks within a given level of linguistic representation. Future work is needed to further develop this perspective on production in more detail.

The Now-or-Never bottleneck, and the processing consequence that follow from it, applies across modalities. Just-in-Time mechanisms of motor planning will be used whether the language output is speech or sign. Similarly, Chunk-and-Pass processing will be required to deal with the onslaught of linguistic material, whether that material is spoken or signed. However, as Emmorey points out, the detailed implications of the Now-or-Never bottleneck may differ between modalities. She notes that the speed of the speech articulators, by contrast with manual gestures, contributes to a rapid serial information transmission strategy being adopted for speech, while greater parallelism is used in signed communication. So, for example, she points out that while spoken words consist of a sequence of phonemes, signed words typically correspond to multiple sign elements (spatial locations and temporally-defined movements). Similarly, she notes that spoken languages deploy affixes temporally before or after the modified item, whereas morphology is usually signaled simultaneously in signed languages. We suggest that differences in sequential learning abilities in the auditory and visual domains may also be important: the perceptual system readily finds sequential structure in auditory material in comparison with visual material (Conway & Christiansen, 2005, 2009; Frost, Armstrong, Siegelman & Christiansen, 2015); conversely, the visual modality readily creates visual ‘Gestalts’ to encode simultaneously presented movements in one or more effectors (compare Bregman, 1994; Wagemans, Elder, Kubovy, Palmer, Peterson, Singh & von der Heydt, 2012—see also Dumitru).

---

1 Chacón et al. contend that “early observations about speech errors indicated that exchange errors readily cross phrasal and clausal boundaries (Garrett, 1980)” (p. 000). A careful reading of Garrett (1980), however, shows that most exchange errors tend to occur within phrases, as would be expected from our perspective.
We have presented Chunk-and-Pass processing as a general solution to the constraints imposed by the Now-or-Never bottleneck. We appreciate the call for proposals concerning how such a framework might elaborated, for example, with respect to the nature of discourse representations (Chacón et al.), developmental underpinnings (Maier & Baldwin), and the nature of processing and representational levels used in Chunk-and-Pass processing (Levinson). In this regard, we are encouraged by the detailed examples provided by O’Grady, illustrating how an account of this kind can be elaborated to deal with linguistically complex phenomena such as wanna contraction, and his more detailed processing-based explanations of central linguistic phenomena including binding and quantification across languages (O’Grady, 2013, 2015).

R3.1 Chunk-and-Pass processing and semantic interpretation

Several commentators (e.g., Chacón et al; Ferreira & Christianson; Frank & Fitz; Honey et al) rightly stressed that a Chunk-and-Pass model of comprehension must integrate current input with past input to produce a semantic interpretation that can interface with general knowledge. The final stages of such interpretation therefore has to do more than merely chunk linguistic input: inferential processes will be required to resolve anaphora and other referring expressions (Garnham & Oakhill, 1985), to bridge between current input and prior linguistic and non-linguistic context (Clark, 1975), and to update beliefs about the speaker’s intentions (e.g., Levinson, 2000) and about the environment (Gärdenfors & Rott, 1995). We argue, though, that the Now-or-Never bottleneck implies that processes of semantic and pragmatic interpretation and belief revision must occur right away, or the opportunity for such interpretation is lost: that is, belief updating, as well as semantic interpretation narrowly construed, is incremental.

The phenomenon of rapid semantic analysis and belief updating is exemplified, for example, in the celebrated demonstration that so-called “close” shadowers (i.e., people able to repeat speech input at a latency of 250-300 ms or even less) are sensitive not only to syntactic structure, but also to semantic interpretation (Marslen-Wilson, 1987). Or consider a very different paradigm, in which a potentially baffling paragraph of text is read either with or without an explanatory title or context (Bransford & Johnson, 1972). In the absence of the explanatory context, memory for the passage is poor. This means that, even if the clarifying context is provided later, the cognitive system is unable to make much sense of the passage in retrospect. Unless it is understood at the time, the details will be too poorly remembered to be reinterpreted successfully. Potter offers a possible framework for such interpretations in terms of what she calls “conceptual short term memory” (CSTM): activations of long-term memory associated with active stimuli and thoughts (Potter, 2012). Importantly, she notes that “rich but unselective associations arise quickly but last only long enough for selective pattern recognition—chunking in CC’s terms.” Thus, CSTM may allow the rapid integration of conceptual information, influenced by task demands and goals, which will facilitate incremental interpretation through Chunk-and-Pass processing. It also enables the building of the kinds of on-line semantic and discourse-related representations called
for by Chacón et al. CSTM may further provide a non-syntactic basis for the successful processing of non-local dependencies (an issue raised by Bever et al.; Healy et al; Levinson).

As noted by Ferreira & Christianson, however, the resulting interpretations may often be rather shallow and underspecified (e.g., Ferreira et al., 2002), with the depth and focus of such “good-enough” representations being affected by task demands (Swets et al., 2008). This can lead to systematic misinterpretations, such as when participants in a study by Christianson, Hollingworth, Halliwell and Ferreira (2001) tended to derive the incorrect interpretation that Mary bathed the baby from the temporally ambiguous sentence While Mary bathed the baby played in the crib. The difficulty of backtracking appears to be a key contributing factor in such misinterpretations because the language system has limited opportunity for going back to correctly reinterpret previous input (Slattery, Sturt, Christianson, Yoshida & Ferreira, 2013).

Our formulation of Chunk and-Pass processing emphasized the importance both of bottom-up and top-down processes. Indeed, we stressed that the pressure rapidly to chunk locally ambiguous speech input provides a powerful reason to harness the full range of relevant informational sources as rapidly as possible. Integrating these sources of information will best predict what input is likely, so that it can be chunked and passed to higher levels of representation as quickly as possible. Parallel models of word recognition (e.g., Marslen-Wilson, 1987; McClelland & Elman, 1986) nicely exemplify this viewpoint: acoustic, lexical, semantic and pragmatic information is brought to bear in real time in order to identify words rapidly, and indeed, the ‘recognition point’ for a word is thereby often reached well before the end of the word. We are therefore highly sympathetic to the call from some commentators to highlight the importance of top-down processing (Dumitru; Healey et al; MacDonald; Potter). Note, though, that top-down expectations from prior context or world knowledge may in some cases also produce misinterpretations, as when participants misinterprets the sentence The man bit the dog as if it was the dog that did the biting (Ferreira, 2003—see also Potter). In such cases, higher-level expectations can run ahead of the linguistic input (as emphasized by Dumitru), potentially leading unanticipated linguistic input to be misinterpreted.

R3.2 The importance of dialogue

Since the turn of the millennium, researchers have become increasingly aware of how viewing language processing in the context of dialogue, rather than considering the isolated production and comprehension of utterances, can have profound implications for the psychology of language (e.g., Pickering & Garrod, 2004). We therefore agree with the various commentators who emphasize the centrality of dialogue in assessing the implications of Chunk-and-Pass processing (Badets; Baggio & Vicario; Healey et al.; Kempson et al.; Levinson)

Kempson et al. and Levinson note the theoretical challenges arising from real-world dialogue in which there is often rapid turn-taking, in which partners may, for example, complete each other’s
sentences. This possibility seems compatible with the idea that production and comprehension processes are closely intertwined (see Pickering & Garrod, 2007, 2013, for reviews). For example, if comprehension involves an ‘analysis-by-synthesis’ reconstruction of the process by which the utterance was produced, then the comprehension process itself creates a representation which can be used to continue the sentence. This is particularly natural within the present framework: the same inventory of chunks can be deployed both by comprehension and production processes. Indeed, a single-system model for processing and producing language in the spirit of the Chunk-and-Pass framework is exemplified in a recent computational model (Chater, McCauley & Christiansen, submitted; McCauley & Christiansen, 2013).

The remarkable speed and rapid turn-taking of interactive dialogue (Levinson) presents a considerable challenge to the cognitive system—although the fact that we are able to understand time-compressed speech which is several times faster than the normal speech-rate, strongly suggests that the limiting factor is rate of articulation rather than comprehension (Pallier, Sebastian-Gallés, Dupoux, Christophe & Mehler, 1998). As Levinson points out, the ability of participants to turn-take with latencies of a fraction of a second implies that significant speech planning has occurred before the other speaker has finished; and prediction is, of course, required to plan an appropriate response before a partner’s utterance is completed. We suggest, though, that on-line prediction and incremental interpretation are required, even when such constraints are relaxed: unless the speech signal is not recoded right away, it will be obliterated by interference from later material. Thus, the ability to anticipate later material at higher levels of representation (e.g., at a discourse level) requires rapid on-line analysis at lower levels of representations, so that the output of such analysis can be fed into predictive mechanisms.

Levinson points out that dialogue can include repetitions that span over fairly long stretches of material, e.g., repeating a query after some intervening comments. Note that this does not provide a problem for the present approach, as long as that material has been recoded into a more abstract form. The Now-or-Never bottleneck implies that maintaining a representation of an acoustic stream, a string of arbitrary acoustic instructions, or a string of phonemes, will be impossible; but such information can be recalled, at least with much greater accuracy, when encoded into a hierarchy of larger units. Hence this type of example is entirely compatible with the Now-or-Never bottleneck.

Rapid interactive dialogue often goes wrong: we continually self-correct, or correct each other (Healey et al.). The ability to do this provides further evidence for an incremental Chunk-and-Pass model of language comprehension—the incrementally created chunked representation can then be revised and reformulated, as appropriate, by making fairly local modifications to the chunk structure. We agree with Healey et al., Kempson et al. and Baggio & Vicario that the ability to switch and repair is a source of strong constraints on theories of processing and, to the extent that the structure of processing matches linguistic structure (Pulman, 1985), by extension to syntactic theory, arguably favoring approaches such as dynamic syntax (Healey et al., Kempson et al.) and construction grammar (O’Grady).
R3.3.3 Meeting the computational challenges of natural dialogue

Huyck and Gómez-Rodríguez highlight the parallel between our framework and the computational solutions used by engineers to implement real-time natural language processing (NLP). Of particular interest is the observation that such artificial systems are not subject to whatever hardware limitations the human brain may be working under but nonetheless end up employing the same solution. One possibility is that the limitations of the brain are actually shaped by the problem, as Lotem et al. suggest. Another possibility is that NLP systems are dealing with human language, which is adapted to the Now-or-Never bottleneck (as discussed further below), and therefore has a very local structure. Artificial NLP system must process language that embodies these human constraints—and, to replicate natural human conversational interaction successfully, they may need to embody those very same constraints. Importantly, Gómez-Rodríguez argues in favor of the latter, because computers are not limited by memory to the same degree as humans. But these systems face the same the same problems as humans when interacting with another person: language needs to be processed here-and-now so that responses can be made within a reasonably short amount of time (e.g., there are about 200 msec between turns in human conversation; Stivers et al., 2009). For example, so-called chatbots (e.g., Wallace, 2005) receive human language input (in text or voice) and produce language output (in text or synthesized voice) in real time. As no one is willing to wait even a few seconds for a response, and because we expect responses even to our half-formed, fragmentary utterances, these chatbots need to process language in the here-and-now, just like people. The strategies they employ to do this are revealing: as Gómez-Rodríguez notes, these artificial systems essentially implement the same Chunk-and-Pass processing solutions that we discussed in our target article: incremental processing, multiple levels of linguistic structure, predictive language processing, acquisition as learning to process, local learning, and on-line learning to predict. We see this convergence as further evidence in favor of the feasibility of Chunk-and-Pass processing as a solution to the pressures from the Now-or-Never bottleneck.

R3.4 Chunk-and-Pass in non-linguistic domains?

If, as we have argued, the Now-or-Never bottleneck is a domain-general constraint on memory, then we should expect Chunk-and-Pass processing to apply not just to language comprehension but also to a wide range of perceptual domains. Similarly, it seems likely that the principles of Just-in-Time production may be extended beyond speech production to action planning and motor control in general (MacDonald; Maier & Baldwin). Indeed, as we noted in our target article, planning one’s own actions and perceiving the actions of others appear to involve the creation of multi-level representational hierarchies, and we conjecture that Chunk-and-Pass and Just-in-Time processes will operate in these domains (Botvinick, 2008; MacKay, 1987).
In the target article, we speculated that music might be a domain in which Chunk-and-Pass and Just-in-Time mechanisms might be required to process a highly complex and hierarchically structured auditory sequence, of comparable complexity to human language. **Lakshmanan & Graham** appear skeptical, apparently on the grounds that music and language differ in a number of regards (e.g., music does not have a semantics; or music does not involve turn-taking—although improvised styles of music including jazz and Indian classical music do frequently involve rapid turn-taking between players). But these concerns seem beside the point when considering the key question at issue: music and language appear to share a hierarchical organization, and both can be processed highly effectively despite the severe pressure of the Now-or-Never bottleneck, and far better than humans can process unstructured sequences of sounds (Warren, Obusek, Farmer & Warren, 1969). We therefore believe that the Chunk-and-Pass framework might fruitfully be applied in future studies of music and other aspects of perception and action.

**R4. Implications for language acquisition, evolution and structure.**

The Now-or-Never bottleneck and its processing consequences (Chunk-and-Pass comprehension and Just-in-Time production) have, we argue, implications for how language is acquired, how it changes and evolves over time, and for the structure of language. The commentators have raised important issues in each of these domains.

**R4.1 Implications for language acquisition**

In our target article, we argued that the Now-or-Never bottleneck implies that language learning is on-line: learning must occur as processing unfolds, or the linguistic material will be obliterated by later input, and learning will not be possible. For parameter-based models of language, this can be difficult—learning seems to require surveying a large corpus of linguistic input to ‘check’ the appropriateness of parameter settings. But if learning must occur on-line, without the ability to retain, and review, a large verbatim corpus, then parameter setting is difficult (witness the difficulties of making ‘trigger’ models in the Principles and Parameters tradition [Gibson & Wexler, 1994] learn successfully). An item-based model of language acquisition provides an alternative conception of on-line learning—new constructions can be added to the learner’s model of the language one-by-one. Such item-based models also fit well with empirical evidence on child language acquisition (e.g., Tomasello, 2003), as well as with item-based models of linguistic structure, such as construction grammar (e.g., Goldberg, 2006).

**Wang & Mintz** characterize this view as a “model-free” approach to language, contrasting it with a “model-based” perspective incorporating linguistic constraints. They suggest that because our approach involves domain-general learning, it is unable to capture many of the constraints on linguistic structure (such as the apparent sensitivity to structure, rather than linear order, in
question formation). This suggestion incorrectly presupposes that domain-general learning necessarily has to be constraint-free. All too often it is implicitly assumed that either language acquisition is guided by (presumably innate) linguistic constraints or there can be no constraints at all. But this is, of course, a false dichotomy. Indeed, we have argued elsewhere (e.g., Christiansen & Chater, 2008, 2016) that there are substantial constraints on language, deriving from a wide variety of perceptual, communicative and cognitive factors (we discuss this point further below). Of these constraints, the Now-or-Ne ver bottleneck is of particular importance but it is not the only one—and so we agree with Wang & Mintz that many additional constraints will shape both language itself and our ability to acquire it. The stronger the confluence of multiple cognitive and other biases that shape language, the easier language will be to learn, because each generation of learners simply have to ‘follow in the footsteps’ of past learners. Language has been shaped by many generations of cultural evolution to fit with our learning and processing biases as well as possible. Thus, considering language as culturally evolved to be easy to learn and process helps explain why language is learned so readily (e.g., Chater & Christiansen, 2010). This viewpoint fits nicely with the iterative learning studies (Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009) described by Lewis & Frank, and their emphasis on language as emerging from the interaction of cognitive and communicative pressures.

Compatible with this viewpoint, and in contrast to Lakshmanan & Graham’s suggestion of acquisition guided by (unspecified) “innate grammatical mechanisms,” Kelly, Wigglesworth, Nordlinger and Blythe (2014), in their survey of the acquisition of polysynthetic languages, highlight the importance of several properties of the input in explaining children’s patterns of acquisition. For example, children learning Quiché Mayan and Mohawk initially produce the most perceptually prominent units of speech, and such perceptual salience also appears to play a role in the acquisition of Navajo, Inuktitut, Quechua, and Tzeltal (Lakshmanan & Graham’s stipulations notwithstanding). Another property of the input—frequency—has been shown by

\[\text{Wang & Mintz}\] seem to have misunderstood the aim of the modeling by Reali and Christiansen (2005). Their point was not to provide a full-fledged model of so-called auxiliary fronting in complex yes/no questions (such as \textit{Is the dog that is on the chair black?}) but rather to demonstrate that the input to young children provided sufficient statistical information for them to distinguish between grammatical and ungrammatical forms of such sentences. Kam, Stoyreshka, Tornyova, Fodor and Sakas (2008) noted some limitations of the simplest bigram model used by Reali and Christiansen, but failed to address the fact that not only did the model fit the results from the classic study by Crain and Nakayama (1987) but also correctly predicted that children should make fewer errors involving high-frequency word chunks compared to low-frequency chunks in a subsequent question elicitation study (Ambridge, Rowland, & Pine, 2008; see Reali & Christiansen, 2009). For example, higher rates of auxiliary-doubling errors occur for questions where such errors involved high-frequency word category combinations (e.g., more errors such as \text{*Is the boy who is washing the elephant is tired?} than \text{*Are the boys who are washing the elephant are tired?}). Most important for current purposes is the fact that Reali and Christiansen—in line with our account of Chunk-and-Pass processing—do not assume that distributional information is all there is to language acquisition: “Young learners are likely to rely on many additional sources of information (e.g., semantic, phonological, prosodic) to be able to infer different aspects of the structure of the target language.” (Reali & Christiansen, 2009: 1024).
Xanthos et al. (2012) to be key to the acquisition of complex morphology across a typologically diverse set of languages: French, Dutch, German (weakly inflecting languages); Russian, Croatian and Greek (strongly inflecting languages); and Turkish, Finnish and Yucatec Maya (agglutinating languages). Using corpus analyses, Xanthos et al. (2012) found that the frequency of different morphological patterns predicted the speed of acquisition of morphology, consistent with usage-based suggestions regarding the importance of variation in the input for learning complex patterns in language (e.g., Brodsky, Waterfall & Edelman, 2007) as well as for distributional learning more generally (e.g., Gómez, 2002).

Lakshmanan & Graham suggest that “without independent grammatical mechanisms” Chunk-and-Pass processing cannot explain children’s acquisition of “free word order” languages such as Tamil. However, a recent computational model by McCauley and Christiansen (2014, 2015c) casts doubt on this claim. This chunk-based learner (CBL) implements Chunk-and-Pass processing at the word level, using simple statistical computations to build up an inventory of chunks consisting of one or more words, when exposed to child-directed speech from a typologically broad set of 29 old-world languages, including Tamil. Importantly, the model works entirely incrementally using on-line learning, as required by the Now-or-Never bottleneck. Following our idea, expressed in the target article, that acquisition involves learning how to process language, CBL gradually learns simplified versions of both comprehension and production. “Comprehension” consists of the chunking of natural child-directed speech, presented to the model word-by-word (essentially, a variation of “shallow parsing”, in line with evidence for the relatively underspecified nature of child and adult language comprehension; e.g., Frank & Bod, 2011; Gertner & Fisher, 2012; Sanford & Sturt, 2002—see also Ferreira & Christianson). In “production,” the task of CBL is to recreate the child utterances encountered in the corpus, given the inventory of chunks learned thus far in the acquisition process. When exposed to a corpus of Tamil child-directed speech, the model was able to use its inventory of chunks to successfully produce a large proportion of the child utterances in that corpus in the absence of “independent grammatical mechanisms”. Indeed, CBL performed as well on Tamil as it did on Mandarin and English. Although not a definitive proof, the CBL simulations do suggest that Chunk-and-Pass processing may be more powerful than a priori speculations might suggest. This underscores the importance of implementing theoretical accounts

3 Endress & Katzir (see also Wang & Mintz) raise a common concern relating to usage-based models: that the sparseness of the input will prevent them from being able to process novel word sequences that are grammatical but not predictable (such as Evil unicorns devour xylophones). Reali, Dale and Christiansen (2005) addressed this challenge head-on, showing in a statistical learning experiment that human participants become sufficiently sensitive to the regularities of training examples to recognize novel sequences whose bigram transitions are absent in training. They subsequently showed that a simple recurrent network (Elman, 1990) could correctly process sequences that contain null-probability bigram information by relying on distributional regularities in the training corpus. Thus, in contrast to the claims of Endress & Katzir, distributional learning appears to be sufficiently powerful to deal with unpredictable but grammatical sequences such as Chomsky’s (1957) famous sentence Colorless green ideas sleep furiously (see also Allen & Seidenberg, 1999).
computationally—whether these accounts are usage-based or rely on innate grammatical mechanisms—in order to determine the degree to which they account for actual linguistic behavior.

Maier & Baldwin raise important questions about how item-based acquisition gets off the ground: for example, what principles can the learner use to establish the basic units from which structures can be built? One possible answer is that information-theoretic properties of the sequence (e.g., points of unusually low predictability) may provide clues to chunk boundaries. A simplified version of this approach is employed by the CBL model (McCauley & Christiansen, 2014, 2015c), which uses dips in backward transitional probabilities (which infants track; cf. Pelucchi, Hay, & Saffran, 2009) to chunk words together. Another approach might discover chunks by way of undersegmentation, essentially treating intonational units as preliminary chunks. The PUDDLE model of word segmentation (Monaghan & Christiansen, 2010) adopts this method and is able to build a vocabulary by using shorter chunks to split up larger chunks. For example, the model is able to use the frequent occurrence of a child’s name in isolation to segment larger utterances in which that name also appears, mirroring the kind of developmental data (e.g., Bortfeld, Morgan, Golinkoff & Rathbun, 2005) that Maier & Baldwin mention. As discussed in McCauley, Monaghan and Christiansen (2015), the two ways of discovering chunks in CBL and PUDDLE likely occur side-by-side in development, possibly alongside other mechanisms. Future research is needed to fully understand the interplay between these different mechanisms and their specific characteristics. Fortunately, as Maier & Baldwin point out, there is considerable empirical evidence that can potentially help constrain models of initial chunk formation (for reviews, see e.g., Arnon & Christiansen, submitted; Werker, Yeung & Yoshida, 2012).

R4.2 Implications for language change and language evolution

Item-based models of language processing and acquisition imply an item-based model of language change. So, assuming that items can be identified with constructions at various levels of abstraction (e.g., from individual lexical items, all the way to constructions determining, for example, canonical word order), then the structure of the language, both within a person, and across individuals, will change construction-by-construction, rather than through the flipping of an abstract parameter which may have diverse and widespread implications (e.g., Lightfoot, 1991). Note, though, that more abstract constructions may be relevant to a large number of sentences of the language. So the fact that the language changes one construction at a time does not imply, for example, that it changes one sentence at a time.

We see language change at the level of language community as an accumulation of changes within the set of constructions acquired by the members of the community. And we view the evolution of language as nothing more than language change writ large. In particular, this implies
that we see the evolution of language as a result of processes of cultural evolution over long periods of human history constrained by communicative goals, as well as our cognitive and neural machinery, rather than resulting from the biological evolution of a language faculty through processes of natural selection or some other mechanism. In short, language is shaped by the brain, rather than the brain being shaped by language (Chater & Christiansen, 2010; Chater, Reali & Christiansen, 2009; Christiansen & Chater, 2008). In the target article, we have aimed to expand on this perspective by exploring how specific properties of language, such as its highly local structure, the existence of duality of patterning, and so on, might arise given the powerful constraint imposed by the Now-or-Never bottleneck.

In this light, Endress & Katzir’s concern that we may be conflating the cultural and biological evolution of language can be set aside; we explicitly reject the idea that there is any substantive biological evolution of language (any more than there has been substantive biological evolution of any other cultural form, whether writing, mathematics, music or chess) although, of course, there will be an interesting biological evolutionary story to tell about the cognitive and neural precursors upon which language has been built. Similarly, Lotem et al.’s worry that we have forgotten about biological evolution is also misplaced. The ‘fit’ between language and language-users arises because language is a cultural product that is shaped around us (and our memory limitations), rather than a fixed and exogenously-given system to which the brain must adapt. Indeed, Our perspective aligns with Charles Darwin’s suggestion that the cultural evolution of language can be viewed as analogous to biological evolution through natural selection. As early as in The Descent of Man, Darwin discussed the cultural evolution of linguistic forms in light of biological adaptation: “The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same.” (Darwin, 1871, p. 59).

One of the great challenges of evolution by natural selection is to explain how biological organisms can increase in complexity. Darwin’s answer was that such complexity may be favored if it increases the number of off-spring at the next generation—i.e., if it improves ‘fitness.’ A parallel challenge arises for explaining the presumed increase in complexity of human languages, from, we may assume, initially limited systems of signed or vocal communication, to the huge richness in phonology, morphology, vocabulary, and syntax, of contemporary natural languages, an issue raised by Behme. Indeed, gradual increases in complexity can happen relatively quickly, as indicated by the fact that children can “outperform” the adults from whom they learn language (Singleton & Newport, 2004), and the incremental incorporation of new linguistic structures into emergent languages such as the Nicaraguan Sign Language (Senghas, Kita & Özyürek, 2004) or the Al-Sayyid Bedouin Sign Language (Sandler, 2012). The pressure for such increases in complexity seems clear: the drive to communicate. While some theorists have argued that the language is not primarily ‘designed’ for communication, but rather for thought (e.g., Chomsky, 2010), we suggest that the social importance of communication underlies the continual generation of new linguistic items, and the
recombination of existing items in creative new ways. Of course, such forms are then subject to the forces of simplification and erosion when they are transmitted across generations of speakers—the forces described by theories of grammaticalization. The picture of language as attempting to maximize communication richness, in the face of memory constraints, is elegantly outlined by Lewis & Frank.

Bergmann et al. note that language change can be affected by the nature of the language community. For example, the presence of a large number of second language speakers (and the properties of their first language) will affect how the new language is processed and transmitted. After all, the Chunk-and-Pass machinery built for a first language will typically be recruited to process a second language, resulting in non-native patterns of chunking. Preliminary support for this perspective comes from analyses of the productions of first (L1) and second language (L2) learners using the earlier mentioned CBL model (McCauley & Christiansen, 2014, 2015c). McCauley & Christiansen (2015a) used the CBL model to compare the “chunkedness” of productions by native Italian speakers learning English or German, when compared to either child or adult native speakers of English and German. The results showed that, compared to those of the L2 speakers, the productions of the native speakers—whether children or adults—were considerably more chunked as measured by repeated multiword sequences. The inability for L2 speakers to chunk incoming input in a native-like way is likely to negatively influence their mastery of fundamental regularities such as morphology and case (Arnon & Christiansen, submitted). In languages with a preponderance of non-native speakers, the L2 learners may exert a greater pressure to regularize and otherwise simplify the language, as Bergmann et al. point out. Thus, the impact of the Now-or-Never bottleneck and the specifics of Chunk-and-Pass processing will vary to some degree based on individual experiences with particular languages.

Viewing language change as operating construction-by-construction does not necessarily rule out the possibility of abrupt change, as we noted above—modifying a single abstract construction (e.g., a ditransitive construction, Subj V Obj1 Obj2) may have far-reaching consequences. Hence, we can set aside Endress & Katzir’s contention that our approach is inconsistent with a part of the literature, which they suggest reports that language change is abrupt and substantial.

The question of whether language change provides evidence for modifications of “deep” linguistic principles or operates construction-by-construction is by no means settled in the literature, as is the question of whether macroscopic linguistic change in a community over historical time is actually abrupt at the level of individual speakers (e.g., Hopper & Traugott, 1993; Lightfoot, 1991; Wang, 1977—an issue parallel to the gradualist vs. punctate equilibrium controversy in biological evolutionary theory, e.g., Dawkins, 1986; Eldredge & Gould, 1972 ). If compelling evidence could be found suggesting that language change involves modification of highly abstract linguistic principles not embedded in a single construction, then this would contradict the item-based model that we see as following from the Now-or-Never bottleneck. But we do not believe that the literature provides such evidence.
R4.3 Implications for language structure

Commentators provide a wide range of viewpoints concerning the relationship between the present account and language structure. A particular concern is how far the Now-or-Never bottleneck is able to capture so-called language universals. We stress that we see broad patterns across languages, whether exception-less or merely statistical universals, as arising from the interaction of a multitude of constraints, including perceptual and cognitive factors, communicative pressures, the structure of thought, and so on (Christiansen & Chater, 2008). Moreover, the trajectory of change observed for a particular language will also be determined by a range of cultural and historical forces, including sociolinguistic factors, language contact, etc. In view of the interaction of this broad range of factors, it may be unlikely that many aspects of language are strictly universal, and indeed, human languages do seem to exhibit spectacular variety, including on such basic matters as the nature and number of syntactic categories. Yet even if strict language universals are, to some degree at least, a myth (Evans and Levinson, 2009), we should nonetheless expect that language will be shaped, in part, by cognitive constraints, such as the Now-or-Never bottleneck.

In this light, concerns that the Now-or-Never bottleneck does not provide an account of all putatively universal features of language (Bever et al.; Chacón et al.; Endress & Katzir) can be set aside. Explaining the cross-linguistic patterns they mention using the multiple constraints discussed above is likely to be a valuable direction for future research. Indeed, we would argue that the Now-or-Never bottleneck is a specific and concrete example of the type of cognitive constraint that Bever et al. believe to underlie universal or near-universal features of language.

Kempson et al. argue that the Now-or Never Bottleneck and its implications have interesting links with formal theories of grammar, such as dynamic syntax, in which there is a close relationship between grammatical structure and processing operations. Similarly O’Grady suggests that the processing bottleneck is manifested differently in different languages. We agree insofar as memory limitations arise from the interaction of the cognitive system with the statistical structure of the language being learned. O’Grady’s specific proposals here and elsewhere (2013, 2015) provide a promising direction for the development of a detailed, and cross-linguistically valid, analysis, linking structure and processing in a way that is consistent with the Now-or-Never bottleneck.

One property mentioned by several commentators as being a widespread (Chacón et al., Levinson; MacDonald) if not universal property of language (Bever et al.) is the existence of non-local dependencies. We have provided a broad account of complex recursive structures incorporating long-distance dependencies elsewhere (Christiansen & Chater, 2015, 2016). Here, we briefly discuss an often-cited example of long-distance dependencies in the form of center-embedding as exemplified in (1) and (2), where the subscripts indicate subject-noun/verb relationships:
(1) The chef who the waiter appreciated admired the musicians.

(2) The chef who the waiter who the busboy offended admired the musicians.

Whereas (1) is easy to comprehend, (2) creates problems for most people (e.g., Blaubergs & Braine, 1974; Hakes, Evans, & Brannon, 1976; Hamilton & Deese, 1971; Wang, 1970). This problem with multiple long-distance dependencies is not unique to English but has also been observed for center-embedded constructions in French (Peterfalvi & Locatelli, 1971), German (Bach, Brown, & Marslen-Wilson, 1986), Spanish (Hoover, 1992) Hebrew (Schlesinger, 1975), Japanese (Uehara & Bradley, 1996) and Korean (Hagstrom & Rhee, 1997). Indeed, corpus analyses of Danish, English, Finnish, French, German, Latin, and Swedish (Karlsson, 2007) indicate that doubly center-embedded sentences such as (2) are practically absent from spoken language. Evidence from sequence learning suggests that the problems with multiple center-embeddings do not derive from semantic or referential complications but rather are due to basic memory limitations for sequential information (de Vries, Geukes, Zwitserlood, Petersson & Christiansen, 2012), as discussed in the target article. These memory limitations may even result in the kind of “illusion of grammaticality” noted by Chacón et al. as when the second verb in (2) is removed to yield the sentence in (3), which to many people seems quite acceptable and even comprehensible (e.g., Christiansen & MacDonald, 2009; Gibson & Thomas, 1999; Vasishth, Suckow, Lewis & Kern, 2010):

(3) The chef who the waiter who the busboy offended admired the musicians.

However, these memory limitations interact with the statistics of the language being used (as discussed previously) such that the above “missing verb” effect can be observed in French (Gimenes, Rigalleau & Gaonac’h, 2009) but not in German (Vasishth et al., 2010) or Dutch (Frank, Trompenaars & Vasishth, in press). Because verb-final constructions are common in German and Dutch, requiring the listener to track dependency relations over a relatively long distance, substantial prior experience with these constructions likely has resulted in language-specific processing improvements (see also Engelmann & Vasishth, 2009; Frank et al., in press, for similar perspectives). Nonetheless, in some cases the missing verb effect may appear even in German, under conditions of high processing load (Trotzke, Bader & Frazier, 2013). We would expect that other non-local dependencies (e.g., as noted by Bever et al., Chacón et al., Levinson, MacDonald) would be amenable to similar types of explanation within the framework of Chunk-and-Pass processing (as also noted by Kempson et al. and O’Grady).

R5. Conclusions and Future Directions

Our target article highlights a fundamental constraint imposed by memory interference on the processing and production of sequential material, and in particular, on language. Dealing with this Now-or-Never bottleneck requires, we argue, chunking and recoding incoming material as
rapidly as possible, at a hierarchy of representational levels (this is Chunk-and-Pass processing). Similarly, it requires specifying the representations involved in producing language just before they are used (this is Just-in-Time production). These proposals themselves have, we suggest, a variety of implications for language structure (e.g., that such structure is typically highly local), for acquisition, and for language change and evolution (e.g., that language changes construction-by-construction both within individuals during learning, and over generations within entire language communities). The commentaries on our article have raised important issues of clarification (e.g., differentiating the present proposals from bottom-up, syntax-driven models such as the Sausage Machine, Frazier & Fodor, 1978); have clarified important links with prior models and empirical results (e.g., the link with ‘good enough’ parsing, Ferreira & Christianson); and outlined supporting evidence (e.g., from the time-course of neural activity involved in language processing, e.g., Honey et al.) and pointed out ways in which the approach can be deepened and made more linguistically concrete (O’Grady). One commentator fears that our proposals may be unfalsifiable (Levinson); others suspect that our approach may actually be falsified by known features of language structure (Bever et al), processing (MacDonald), acquisition (Wang & Mintz), or language change (Endress & Katzir). We hope that our target article will persuade readers that memory constraints have substantial implications for understanding many aspects of language; and that our response to commentators makes the case that the many claims flowing from the Now-or-Never bottleneck are compatible with what is known about language (although not always with what is presumed to be case by prior theories). Most importantly, we encourage interested readers to continue the work of the many commentators who provide constructive directions to further explore the nature of the Now-or-Never bottleneck, further elaborate and test the Chunk-and-Pass and Just-in-Time perspectives on language processing, and to help integrate the study of these performance constraints into our understanding of key aspects of language structure, acquisition and evolution (for some steps in this direction, see Christiansen & Chater, 2016).

References


Schlesinger, I. M. (1975). Why a sentence in which a sentence in which a sentence is embedded is difficult. Linguistics, 153, 53-66.


