processing preference fails to outlaw all of the discontinuous structures in language, and where our push-down stack capacities actually reside.

Finally, C&C’s Now-or-Nevert bottleneck theory suggests that details of an utterance cannot be retained in memory when following material overwrites it—only the gist of what was said may persist. But the practice of “other-initiated repair” suggests otherwise—so the following excerpt Sig repeats verbatim what he earlier said, just with extra stress on shoot even though three conversational turns intervene (Schegloff 2007, p. 109):

(1) Sig: Conservatives like to shoot people (and liberals don’t?)
   (2.0)
Dad: Conservatives like what?
   (0.8)
Sig: What?
   (3.0)
Dad: Whadyu say about conservatives? (lack of stress)
   (0.3)
Sig: Conservatives like to shoot people on (hi) liberals don’t?

The fact that we can rerun the phonetics (? = rising intonation, underlining = stress) of utterances shows the existence of other buffers that escape the proposed bottleneck.

Linguistic structure emerges through the interaction of memory constraints and communicative pressures
doi:10.1017/S0140525X15000874, e82

Molly L. Lewis and Michael C. Frank
Department of Psychology, Stanford University, Stanford, CA 94305.
ml@stanford.edu mcfrank@stanford.edu
http://web.stanford.edu/~miller/
http://web.stanford.edu/~mcfrank/

Abstract: If memory constraints were the only limitation on language processing, the best possible language would be one with only one word. But to explain the rich structure of language, we need to posit a second constraint: the pressure to communicate informatively. Many aspects of linguistic structure can be accounted for by appealing to equilibria that result from these two pressures.

Christiansen & Chater (C&C) claim that memory limitations force the cognitive system to process the transient linguistic signal by compressing it. They suggest that this processing pressure influences the ultimate structure of language over the course of language evolution. Taken at face value, this proposal would lead to a degenerate linguistic structure, however. If memory constraints were the only pressure on language, languages would evolve to compress meaning into the simplest possible form—a single word (Horn 1984). But, as the authors point out, natural languages are not of this sort; they are richly structured into lexical and phrasal units of varying length. To account for this variability, we highlight the need to consider the communicative function of language. Communication serves as an important counter-pressure against compression in language processing, not just as a caveat.

Intercutors use language with the goal of communicating information, but they also aim to minimize energetic cost (Zipf 1949). For the speaker, this goal implies minimizing production cost, and for the listener it implies minimizing comprehension cost. Importantly, these processing constraints have opposing cost functions (Horn 1984; Zipf 1949). For a producer, processing is minimized when a form is easy to say, and thus highly compressible. For the comprehender, however, processing is minimized when a form is minimally ambiguous and thus verbose. Compressing information is a useful strategy for a speaker who faces memory constraints, but it is useful only to the extent that the listener can still recover the intended meaning. This view of language use as rational action—minimizing costs while maximizing information transfer—is supported by a rich body of theoretical and empirical work (Clark 1996; Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Grice 1975).

Although C&C argue that compression is the key factor in the emergence of structure, evidence at both the acquisition and evolution timescales suggests language is the product of the interaction between both compression and informativity. At the timescale of acquisition, experimental work suggests the resolution of reference in word learning is the product of communicative inferences (e.g., Baldwin 1991; 1993; Frank et al. 2009; Frank & Goodman 2014). And at the timescale of language evolution, a growing body of work suggests that the forms of words are also equilibria between these two pressures (Lewis & Frank 2014; Mahowald et al. 2012; Plantadosi et al. 2011; Zipf 1936). For example, Plantadosi et al. (2011) found that words that are less predictable in their linguistic context are longer, suggesting that speakers may lengthen words that are surprising in order to increase time for the listener to process.

In addition to linguistic form, these pressures influence the mapping between form and meaning. An equilibrium in the structure of form-meaning mappings is one in which the listener is able to recover the intended meaning, but the speaker does not exert additional effort over-describing. A range of semantic domains reflect this equilibrium (Baddeley & Attewell 2009; Kemp & Regier 2012; Regier et al. 2007), and ambiguity, more generally, has been argued to reflect this communicative tradeoff (Plantadosi et al. 2012). Ambiguity is an equilibrium in cases where the listener can recover the intended meaning from the communicative context. One example is the word “some,” which has a literal meaning of “at least one and possibly all” but can be strengthened pragmatically to mean “at least one but not all” (Horn 1972). Because its meaning is determined through communicative context, its literal semantics can overlap those of its competitor, “all.”

The key challenge associated with this broader proposal—that processing pressures influence linguistic structure—is providing direct evidence for a causal link between these two timescales. This problem is difficult to study in the laboratory because the proposed mechanism takes place over a long timescale and over multiple individual speakers. Furthermore, the presence of a causal link does not entail that phenomena in processing are directly reflected in linguistic structure—rather, entirely new properties may emerge at higher levels of abstraction from the interactions of more fundamental phenomena (Anderson 1972). It may, therefore, not be possible to directly extrapolate from brief communicative interactions observed in the laboratory to properties of linguistic structure.

Several recent pieces of experimental data begin to address this challenge, however. In one study, Fedzechkina et al. (2012) asked speakers to learn an artificial language that arbitrarily distinguished nouns through case-marking. Over learning sessions, speakers developed a system for marking in contexts where meanings were least predictable—a pattern reflected in the case-marking systems of natural language. Other work has used a similar paradigm to reveal the emergence of typologically prevalent patterns in the domains of word order (Culbertson et al. 2012; Culbertson & Newport 2015) and phonology (Wilson 2008).

A particularly promising approach for exploring this causal link is through transmission chains (Kirby et al. 2008; Reali & Griftiths 2009). In a transmission chain, a participant learns and recalls a language, and then the recalled language becomes the learning input for a new learner. By iterating over learners, we can observe how languages change across transmission of learners over the course of language evolution. Kirby et al. (2015) have compared the emergence of linguistic structure in a
The bottleneck may be the solution, not the problem

doi:10.1017/S0140525X15000886, e83

Amon Lotem,a Oren Kolodny,b Joseph Y. Halpern,c and Shimon Edelmana

aDepartment of Zoology, Tel Aviv University, Tel Aviv 6997801, Israel;
bDepartment of Biology, Stanford University, Stanford, CA 94305;
cDepartment of Computer Science, Cornell University, Ithaca, NY 14853; Division of Linguistics and Multilingual Studies, Nanyang Technological University, Singapore 637332; Department of Psychology, Cornell University, Ithaca, NY 14853.

lotem@post.tau.ac.il okolodny@stanford.edu halpern@cs.cornell.edu lucoa@ntu.edu.sg se37@cornell.edu

Abstract: As a highly consequential biological trait, a memory “bottleneck” cannot escape selection pressures. It must therefore co-evolve with other cognitive mechanisms rather than act as an independent constraint. Recent theory and an implemented model of language acquisition suggest that a limit on working memory may evolve to help learning. Furthermore, it need not hamper the use of language for communication.

The target article by Christiansen & Chater (C&C) makes many useful and valid observations about language that we happily endorse. Indeed, several of C&C’s major points appear in our own papers, including the following: (a) the inability of non-chunked, “analog” approaches to language to compete with “digital” combinatorics over chunks (Edelman, 2008b); (b) the centrality of chunking to modeling incremental, memory-constrained language acquisition and generation (Goldstein et al. 2010; Kolodny et al. 2015b) and the possible evolutionary roots of these features of language (Kolodny et al. 2014; 2015a; Lotem & Halpern 2012); (c) the realization that language experience has the form of a graph (Solan et al. 2005; cf. Edelman 2008a, p. 274), corresponding to C&C’s “forest tracks” analogy; and (d) a proposed set of general principles for language acquisition and processing (Goldstein et al. 2010), one of which is essentially identical to C&C’s “Now-or-Ne ver bottleneck.” However, our theory is critically different in its causality structure. Rather than assuming that the memory limit is a fixed constraint to which all other traits must adapt, we view it as an adaptation that evolved to cope with computational challenges. Doing so brings theory in line with standard practice in evolutionary biology, is more consistent with research findings, and raises numerous important research issues. We expand on these points in the following paragraphs.

No biological trait can be simply assumed as a “constraint.” Viewing the Now-or-Ne ver bottleneck as an evolutionary constraint to which language adapts – C&C’s central idea – is unwarranted. In evolutionary theory, biological constraints – as opposed to constraints imposed by physics and chemistry, which are not subject to biological evolution – cannot simply be assumed; they must be understood in terms of trade-offs among selective pressures. Clearly, birds’ wings evolved under aerodynamic constraints rather than vice versa. However, biological traits such as memory are not exempt from evolving. In proposing a bottleneck to which everything else in the system must adapt while the bottleneck itself remains fixed and independent (Fig. 1 in the target article), C&C implicitly assume that it cannot evolve.

To justify this assumption, C&C should have offered evidence of stabilizing selection pressures that act against genetic variants coding for a broader or narrower bottleneck, and thereby affecting cognition and, ultimately, fitness. Alternatively, they might have assumed that the biological mechanisms underlying the memory bottleneck cannot be genetically variable – an odd assumption, which runs counter to substantial evidence in humans of (a) a range of verbal memory decay rates (Mueller & Krawitz 2009), including in particular the longer verbal working memory span in individuals with Asperger’s (Cui et al. 2010); (b) heritable variation in language and in word memory (Stromswold 2001; van Soelen et al. 2011) and in working memory (Blokhland et al. 2011; Vogler et al. 2014); and (c) variation in perceptual memory across species (Lind et al. 2015; Mery et al. 2007). Given that heritable variation in a trait means that it can respond to selection (e.g., F-alone 1981), it is likely that the bottleneck can evolve, and that it is what it is because individuals with longer or shorter verbal working memory had lower biological fitness.

If language is supported by domain-general mechanisms, verbal memory is even less immune to evolution. If the emergence of language constitutes a recent and radical departure from other cognitive phenomena, it is in principle possible that working memory evolved and stabilized prior to and separately from the “increasingly abstract levels of linguistic representation” (sect. 3.2, para. 2) posited by C&C. However, there are good arguments in support of a domain-general view of language (e.g., Chater & Christiansen 2010). In particular, linguistic representations and processes are hardly as modular as C&C assume (Onnis & Spivey 2012). Furthermore, theories of neural reuse (Anderson 2010) point to the massive redeployment of existing mechanisms for new functions, resulting in brain regions coming to be involved in diverse cognitive functions. If circuits that support language continue contributing to nonlinguistic functions (including working memory), a memory bottleneck is not a prior and independent constraint on language, but rather a trait that continues to evolve under multiple selective pressures, which include language.

The bottleneck may be the solution, not the problem. As we have suggested (Goldstein et al. 2010; Lotem & Halpern 2008; 2012; Onnis et al. 2008), a limited working memory may be an adaptation for coping with the computational challenges involved in segmenting and network construction. (Importantly, regardless of whether this specific hypothesis is correct, entertaining such hypotheses is the only way of distinguishing a function from a constraint; cf. Stephens & Krebs 1986, Ch. 10.) A recently implemented model that includes this hypothesis has been tested on tasks involving language, birdsong, and foraging (Kolodny et al. 2014; 2015a; 2015b; Menyhart et al. 2015) The model includes a time window during which natural and meaningful patterns are likely to recur and thus to pass a test for statistical significance, while spurious patterns decay and are forgotten. We stress that rather than acting as a constraint, the duration of the window must co-evolve with the mechanisms influencing the distribution of data so as to increase the effectiveness of memory representations (Lotem & Halpern 2012).

We do agree with C&C regarding some of the consequences of the memory bottleneck, such as the need for online incremental construction of hierarchical representation. Indeed, our model effectively implements what C&C call “Chunk-and-Pass” (Kolodny et al. 2015b). We believe, however, that the ultimate constraint on learning structure (such as that of language) in time and space is not the memory bottleneck in itself, but rather the
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 have 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 (Medeiros 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 the case by prior theories). Most important, 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 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).

NOTES

1. Chacón et al. contend that “early observations about speech errors indicated that exchange errors readily cross phrasal and clausal boundaries (Garrett 1980)” (para. 7). A careful reading of Garrett, however, shows that most exchange errors tend to occur within phrases, as would be expected from our perspective.

2. 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 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 et al. (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 (Ambríuste et al. 2009; 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 *Is the boy who is washing the elephant is tired*? than *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, p. 1024).

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 et al. (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).

References

[The letters “a” and “r” before author’s initials stand for target article and response references, respectively.]


References/Christiansen & Chater: The Now-or-Never bottleneck: A fundamental constraint on language

of the Royal Society B: Biological Sciences 369(1634):20120394. [AMHC]
Waxman, S. R. & Gelman, S. A. (2009) Early word-learning entails reference, not...