Journal of Memory and Language 89 (2016) 244–254

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Language as skill: Intertwining comprehension and production

Nick Chater [a,*], Stewart M. McCauley [b], Morten H. Christiansen [b,c,d]

[a] Behavioural Science Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK
[b] Department of Psychology, Cornell University, Ithaca, NY 14853, USA
[c] Department of Language and Communication, University of Southern Denmark, 5230 Odense M, Denmark
[d] Haskins Laboratories, New Haven, CT 06511, USA

ARTICLE INFO

Article history:
Received 14 October 2014
revision received 4 November 2015
Available online 27 November 2015

Keywords:
Comprehension
Language production
Now-or-never bottleneck
Incremental processing
On-line learning
Skill learning

ABSTRACT

Are comprehension and production a single, integrated skill, or are they separate processes drawing on a shared abstract knowledge of language? We argue that a fundamental constraint on memory, the Now-or-Never bottleneck, implies that language processing is incremental and that language learning occurs on-line. These properties are difficult to reconcile with the 'abstract knowledge' viewpoint, and crucially suggest that language comprehension and production are facets of a unitary skill. This viewpoint is exemplified in the Chunk-Based Learner, a computational acquisition model that processes incrementally and learns on-line. The model both parses and produces language; and implements the idea that language acquisition is nothing more than learning to process. We suggest that the Now-or-Never bottleneck also provides a strong motivation for unified perception–production models in other domains of communication and cognition.

Language as knowledge; language as skill

The ability to comprehend and produce language requires spectacular levels of skill. But is it one skill, or two? Transfer between comprehension and production appears to suggest a unitary system; hearing a language is, it seems, crucially important for speaking that language. After all, children don't simultaneously learn to understand German while producing Mandarin. Indeed, across diverse theoretical perspectives in the language sciences, there is general agreement that there exists an important overlap between the knowledge and processing operations involved in comprehension and production. But there is considerable disagreement about the nature of this relationship.

One viewpoint (e.g., Chomsky, 1965), which has been dominant in many theoretical approaches to language, starts by assuming a strong separation between linguistic competence (i.e., an abstract specification of the speaker/hearer's knowledge of the language) and linguistic performance (the processes by which this abstract competence is deployed in language processing). From this standpoint, the overlap between production and comprehension may reside purely in the shared abstract linguistic competence that is being drawn upon by both comprehension and production processes. But the processes of comprehension and production could, in principle, be completely unrelated. We call this the 'language as knowledge' view.

An opposing viewpoint suggests that no such abstract linguistic competence exists—rather, acquiring language is no more than acquiring the ability to process language. There is no separate representation of the abstract structure of the language (e.g., a grammar) distinct from the mechanisms of language production and comprehension; instead there are simply procedures for language processing (e.g., Kempson, Meyer-Viol, & Gabbay, 2001; O'Grady, 2013). From this point of view, the overlap

* Corresponding author.
E-mail address: Nick.Chater@wbs.ac.uk (N. Chater).

http://dx.doi.org/10.1016/j.jml.2015.11.004
0749-596X/

between comprehension and production requires that the same (or highly overlapping) processes underpin both the comprehension and production of utterances. We call this account the 'language as skill' perspective (see also Christiansen & Chater, 2016).

We will argue, in the next section, *Fundamental memory constraints on skill learning*, that basic limitations on memory strongly favor the language as skill perspective, and hence the assumption that the processing operations of comprehension and production are intimately related. This requires viewing language acquisition as a process of on-line skill learning, where processing operations lay down traces that facilitate further processing—there is no opportunity for inferring abstract general principles of language structure. In light of the limitations we outline, the challenge of building cognitively plausible models of language processing and acquisition is considerable. We take some initial steps toward addressing this challenge in Section 'A unified model of production and comprehension', describing a computational model incorporating incremental processing and on-line learning. Crucially, although comprehension and production are closely integrated in the model, we present new simulation results demonstrating that the model gives rise to the kind of comprehension–production asymmetry often observed in language acquisition (e.g., Fraser, Bellugi, & Brown, 1963). In Section 'Integrated production and comprehension', we then reflect on the broader theoretical issues raised by proposing a unitary model, before drawing brief conclusions.

Fundamental memory constraints on skill learning

On just about any measure, language processing is astonishingly fast. Speaking rates are typically 10–15 phonemes per second, which translates to as much as 150 words per minute (Studdert-Kennedy, 1986). This daunting speed implies that the comprehension system is faced with a relentless onslaught of new input. Similarly, the production system has to generate and execute a stream of articulatory instructions at a remarkable speed. Furthermore, the interleaving of comprehension and production processes is also very fast, as made evident by the rapid turn-taking observed across languages and cultures (e.g., the mean latency between 'turns' is typically about 200 ms, Stivers et al., 2009), the ability to 'shadow' speech within a 250 ms latency or less (Marslen-Wilson, 1973, 1985), and our ability to fluently complete each other's sentences (Clark & Wilkes-Gibbs, 1986).

Our impressive performance in processing language contrasts strikingly with our very limited ability to process sequences of arbitrary auditory or visual stimuli. For example, in a classic study, Warren, Obusek, Farmer, and Warren (1969) found that naïve listeners were unable correctly to recall the order of just four different non-speech sounds, even though each could easily be identified in isolation. More broadly, memory for the temporal order of arbitrary stimuli appears restricted to 4 ± 1 items (Cowan, 2000); and even the identities of the items in a sequence is typically rapidly lost, e.g., when measured by probed or free recall (e.g., Baddeley, 2007). In short, memory is fleeting: unless information is recoded and/or used rapidly, it is subject to severe interference from an onslaught of new material. We call this the Now-or-Never bottleneck (Christiansen & Chater, in press).

To cope with the flow of speech input, it is crucial that phonemes are rapidly recoded into higher-level units—for example, into syllables, lexical items, phrases, and beyond (although the specific hierarchy of linguistic levels is, of course, controversial; and may vary from language to language): we call this Chunk-and-Pass language comprehension (Christiansen & Chater, in press). These more abstract levels correspond both to larger units of speech input (so that the same three- or four-item limit corresponds to a longer stretch of speech), and also will typically lead to less interference with subsequent material. This is because, on just about any theoretical account, the space of lexical items is very much larger than space of phonemes, so that confusability between phonemes will be much greater than confusability between lexical items.

Parallel issues arise in speech production. Here, the constraint is to maintain the shortest possible 'inventory' of material waiting to be generated, to avoid interference between items to be produced, at a given level of representation (e.g., Dell, Burger, & Svec, 1997). And as with comprehension, this, too, can only be done by decoding higher-level representations into lower-level representations in a piecemeal manner. If, for example, an entire message were decoded into a string of phonemes before even beginning to speak, that stream of phonemes would vastly exceed the few items we can accurately hold in memory, and hence would rapidly be lost. The production system must, therefore, decode a higher-level representation into a more detailed lower-level representation when that lower level representation will be used, and not substantially before. Thus, while our speech production system may look ahead several words in advance, those words will only be converted into, say, a phoneme representation, when the word is almost ready to spoken; and the phonemes will in turn only be translated into still more detailed articulatory instructions at the very last moment. The cascade of different levels of speech production thus obeys a principle of 'Just in Time' processing, analogous to the inventory management system pioneered in Japanese manufacturing (Ohno & Mito, 1988). A 'stock' of representations cannot be allowed to accumulate, because such stock is highly 'perishable'—specifically, it will be rapidly interfered with or overwritten by the continual arrival of new material (Christiansen & Chater, in press)[1].

The Now-or-Never Bottleneck has particularly striking implications for language acquisition. If linguistic input can only be retained very briefly before being overwritten, then learning must occur *as language is being processed*.

[1] We want to stress that we are not advocating for so-called 'radical incrementality' in production, in which words are articulated immediately without any planning ahead. Rather, we see production as involving planning a few chunks ahead at every level of linguistic abstraction. Importantly, though, 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, and even longer at the conceptual/discourse level (see Chater & Christiansen, in press, for further discussion).