Developmental Science

Developmental Science 12:3 (2009), pp 365–368 DOI: 10.1111/j.1467-7687.2009.00821.x

SPECIAL SECTION: CORE COMPUTATIONAL PRINCIPLES OF LANGUAGE ACQUISITION: CAN STATISTICAL LEARNING DO THE JOB

Core computational principles of language acquisition: can statistical learning do the job? Introduction to Special Section

Bob McMurray[1] and George Hollich[2]

1. *Department of Psychology and the Delta Center, University of Iowa, Iowa City, USA*
2. *Department of Psychology, Purdue University, West Lafayette, IN, USA*

In the last 15 years, the field of language acquisition has undergone a sea-change. Whereas prior work emphasized the nature of the representational structures and capacities that would support learning (e.g. Chomsky, 1995), a large number of studies are now looking directly at the learning processes that are responsible for acquiring the structure of language (Saffran, 2003; Sebastián-Gallés, 2007 for reviews).

It's about time.

Computational work in the connectionist tradition had long made the argument that there was significant information available in the statistics of the input to support language acquisition, and that simple neurally plausible devices could extract this information to do useful work (for classic examples, see Elman, 1990; Rumelhart & McClelland, 1986; and see Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996, for a review). However, to many psychologists these models remained existence proofs without behavioural demonstrations that such learning was possible. However, in 1996, Saffran, Aslin, Newport and colleagues demonstrated that 8-month-old infants (Saffran, Aslin & Newport, 1996) and adults (Saffran, Newport, Aslin, Tunick & Barrueco, 1997) could learn the between-syllable transition probabilities in a stream of connected speech and use them to segment words, even when there were no bottom-up cues to word boundaries.

Since then, a small cottage industry has grown up in which researchers use these paradigms to demonstrate that infants and adults are capable of learning statistics that might support a range of language processes. Such processes include phonetic categories (Maye, Werker & Gerken, 2002; Maye, Weiss & Aslin, 2008), phonological regularities (Newport & Aslin, 2004), aspects of syntax (Thompson & Newport, 2007; Gómez, 2002; Saffran, 2001), and word/referent mappings (Yu & Smith, 2007). Statistical learning work has even begun to look at learning across multiple levels (Graf Estes, Evans, Alibali & Saffran, 2007), and has shown that these learning processes are not restricted to language (Hunt & Aslin, 2001; Saffran, Johnson, Aslin & Newport, 1999; Fiser & Aslin, 2002, 2005; Kirkham, Slemmer & Johnson, 2002).

One strength of statistical learning as a theoretical paradigm is that it posits discrete, clear learning mechanisms, mechanisms such as distributional statistics over phonetic cues (Maye *et al.*, 2002), transition probabilities across an intervening syllable (Gomez, 2002), and word/referent co-occurrence (Yu & Smith, 2007). These are straightforward to measure in the input (to varying degrees), and can be easy to manipulate experimentally. Moreover, theoretically, they usually operate locally between observable units and do not appear to be pre-tuned to language structure. Thus, in their simplicity, these specific statistical learning mechanisms would seem to be clear examples of concrete, core mechanisms of language development.

However, in this simplicity lies a question. Can these core mechanisms do the job? Behavioural investigations of the sort conducted thus far cannot answer this. Although these studies demonstrate without a doubt that such learning occurs, it is an open question whether it can scale up to the massive problem that is language. For example, can transition probabilities be used to segment all 60,000 words in the average lexicon?

Work such as Yang (2004) suggests that this is non-trivial for some classes of statistics: transitional probabilities across syllables alone may not be sufficient for discovering the majority of words (although other classes of regularities might be). Furthermore, even if statistics alone were sufficient to recognize words, would the resulting representations be sufficient for further developmental achievements (e.g. constructing grammatical categories)? It is possible

Address for correspondence: Bob McMurray, Department of Psychology, E11 SSH, University of Iowa, Iowa City, IA 52240, USA; e-mail: bob-mcmurray@uiowa.edu

that statistical learning and the core mechanisms discussed in this section serve to get the process off the ground, but that other language-specific or social/pragmatic mechanisms take over beyond a certain point.

Thus, the circle turns 'round again. The question of the sufficiency of statistical learning mechanisms is best answered computationally. The only way to know whether transition probabilities are truly sufficient for segmentation is to compute them over a large corpus of language. The only way to know if distributional statistics could be useful for phonetic category learning given the noise in the input is to put these discrete computational mechanisms in this context and see. A computational implementation of these theories (a model) has the potential to answer these questions, particularly when coupled with corpora of real language. However, although connectionist simulations motivated much of the statistical learning approach to begin with, they may not be ideal for this next undertaking.

Although such networks do engage in statistical learning, they do not simply acquire the structure of the input. They acquire it *for a specific purpose* (cf. McMurray, Horst, Toscano & Samuelson, in press). Elman's (1990) simple recurrent network, for example, acquires the structure that is useful for predicting the next element in the sequence, whereas Rumelhart and McClelland's (1986) past-tense model acquires the structure that is useful for determining the past tense of English verbs. Other statistics not related to the task may be ignored. As a tool, then, this is incredibly powerful, allowing learning to be focused by useful behaviour. It also reinforces the notion that a connectionist model is not simply a generic, blank-slate learning device (cf. Elman *et al*., 1996). Moreover, unlike Bayesian approaches, the force driving the acquisition of the statistics is useful behaviour – the statistics are only means to an end.

However, this presents challenges when trying to translate these models into theory. What mechanisms beyond statistical learning do such networks include? What theoretical elements does the task impose on top of this learning? Even the simplest back-propagation networks, for example, include non-linear internal representations and dimensionality reduction (e.g. Elman & Zipser, 1988). Many unsupervised networks include competition (Rumelhart & Zipser; 1986; McMurray & Spivey, 1999). Other networks include a bewildering array of mechanisms (e.g. Nakisa & Plunkett, 1998). Are these related to the stability of the representation, the basic learning process, or the ability to complete the task? This can be difficult to parse out.

Thus, although such models represent an important part of our exploration of the mechanisms of language learning, it is also useful to take a step back and examine each of the candidate mechanisms individually. Put another way, from the perspective of identifying candidate mechanisms of language development, connectionist networks typically don't separate the effects of statistical learning from those of other mechanisms, leaving it unclear what is necessary and sufficient to do the job (although in some cases they could, see McMurray *et al*., in press). In a sense, the minimal grain of analysis is the whole network.

What is needed is an approach in which such mechanisms can be separated, analysed and assessed. Luckily, the simplicity of statistical learning mechanisms (as proposed) makes this a tractable task. Hypothetical sources of information such as co-occurrence statistics, the likelihood of various sorts of transitions, and frequency can be implemented computationally, and this in turn can allow their performance to be evaluated against real language data. If other mechanisms are necessary to turn these statistics into useful behaviour, these can be added systematically and their theoretical role understood more clearly. These other mechanisms might be the use of additional classes of statistics, or they may be something else, such as competition between representations, sensitivity to particular perceptual salience, or error-driven learning.

Such analyses are powerful, whether carried out from a statistical learning perspective or from other perspectives (Yang, 2004). This approach, however, should not be limited to statistical learning – such formal and computational transparency should be a hallmark of all theories of development.

To be clear, we are not saying that learning alone, particularly in this distilled form, is sufficient to account for all of the complexities of development. Development is inherently non-rational, interactive and dynamic (cf. Spencer, Blumberg, McMurray, Robinson, Samuelson & Tomblin, in press). Such elementary computations may be modulated by the social environment, the developmental history, and the cognitive abilities of the child (see also Hirsh-Pasek, Golinkoff & Hollich, 2000; Sebastián-Gallés, 2007). Nonetheless, this approach is of immense theoretical value. It allows us to ask if statistical learning mechanisms can do specific developmental jobs or if more is needed. Ultimately, dynamical systems and/or connectionist approaches may provide a more complete account of development as a whole (e.g. Spencer, Thomas & McClelland, in press; Thelen & Smith, 1994; Elman *et al*., 1996). However, for the theoretical task of isolating and understanding specific mechanisms, simple computational techniques confronted with the complexity of real input offer an important complement.

The papers in the present section bring together work in a variety of domains to ask this basic question: Can statistical learning do the job and what other mechanisms may be involved? McMurray, Toscano and Aslin address this in the domain of speech perception; Hollich and Prince examine attention and source localization; Christiansen, Onnis and Hockema look at the interface of word segmentation and grammatical categories; and Chemla, Mintz, Bernal and Christophe model the beginnings of syntactic structure.

There are many avenues on which to evaluate our basic question. Each paper takes a unique approach. McMurray *et al*., for example, ask if acquiring a set of statistics is sufficient for acquiring a useful cognitive representation (in their case, a phonetic category), and Christiansen *et al.* take a parallel tack, asking about the phonological information that may be involved in grammatical categories. We must also ask if particular learning devices are sufficient for real language, beyond the toy-problems we use as proof-of-concept, and beyond the use of English. In this vein, Chemla *et al.* evaluate their frequent-frames approach to grammatical categorization to French.

In addition, language learning should not be compartmentalized into simply learning phonemes, words, or grammatical categories – learning in one domain has consequences for another. Thus, Christiansen *et al.* ask whether the output of a statistical word segmenter can serve as useful input for acquiring grammatical categories, and Chemla *et al.* evaluate whether categories extracted by the frequent-frames statistic can themselves serve as the elements of further statistical learning – can frequent frames be used recursively?

Finally, statistical learning ultimately must support real behaviour. Thus, a crucial question is whether the statistics alone are sufficient to account for some behaviour or if other processes are required. Here, Hollich and Prince ask whether statistical co-occurrence of low-level auditory and visual signals is sufficient to model infants' moment-by-moment source looking behaviour in an auditory/visual integration task, and McMurray demonstrates how the unfolding of phonetic category learning has consequences for early phonetic discrimination.

The papers in this section thus take the exploration of the explanatory power of statistical learning to the next level. The elegance of statistical learning is that it allows us to quantify our theories of development precisely, and to test their sufficiency against real-world data. Although the four papers in this section examine different sub-domains of language and use different metrics for evaluation, all take this idea seriously and push the field of language development forward in interesting ways. In particular, such work begins the formation of a bridge between statistical *learning* and language *development*.

References

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, **14**, 179–211.

Elman, J., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, **83** (4), 1615–1626.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28** (3), 458–467.

Fiser, J., & Aslin, R.N. (2005). Encoding multi-element scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, **134**, 521–537.

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science*, **13**, 431–436.

Graf Estes, K.M., Evans, J., Alibali, M.W., & Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, **18**, 254–260.

Hirsh-Pasek, K., Golinkoff, R., & Hollich, G. (2000). An emergentist coalition model for word learning: mapping words to objects is a product of the interaction of multiple cues. In R.M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. Woodward, N. Akhtar, M., Tomasello, & G. Hollich (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 136–164). New York, NY: Oxford University Press.

Hunt, R., & Aslin, R.N. (2001). Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, **130** (4), 658–680.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). *Cognition*, **83** (2), B35–B42.

McMurray, B., & Spivey, M. (1999). The categorical perception of consonants: the interaction of learning and processing. *Proceedings of the Chicago Linguistics Society*, **35**, 205–219.

McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (in press). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a new grand theory of development? Connectionism and dynamic systems theory reconsidered*. London: Oxford University Press.

Maye, J., Werker, J.F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, **82**, 101–111.

Maye, J., Weiss, D.J., & Aslin, R.N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, **11**, 122–134.

Nakisa, R., & Plunkett, K. (1998). Evolution of a rapidly learned representation for speech. *Language and Cognitive Processes*, **13** (2&3), 105–127.

Newport, E.L., & Aslin, R.N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, **48**, 127–162.

Rumelhart, D., & McClelland, J. (1986). On learning the past tense of English verbs. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition*, *Volume 2* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D.E., & Zipser, D. (1986) Feature discovery by competitive learning. In D.E. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1* (pp. 151–193). Cambridge, MA: MIT Press.

Saffran, J.R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, **44**, 493–515.

Saffran, J.R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, **12**, 110–114.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A., & Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science*, **8**, 101–105.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, **70**, 27–52.

Sebastián-Gallés, N. (2007) Biased to learn language. *Developmental Science*, **10** (6), 713–718.

Spencer, J., Blumberg, M., McMurray, B., Robinson, S., Samuelson, L., & Tomblin, J.B. (in press) Short arms and talking eggs: why we should no longer abide the nativist–empiricist debate. *Child Development Perspectives.*

Spencer, J., Thomas, M., & McClelland, J. (in press) *Toward a new grand theory of development? Connectionism and dynamic systems theory reconsidered.* London: Oxford University Press.

Thelen, E., & Smith, L.B. (1994) *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, **8** (10), 451–456.

Yu, C., & Smith, L. (2007) .Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, **18**, 414–420.