Cognition 178 (2018) 92–102

Contents lists available at ScienceDirect

Cognition

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

Original Articles

When learning goes beyond statistics: Infants represent visual sequences in terms of chunks

Lauren K. Slone*, Scott P. Johnson

Department of Psychology, University of California, Los Angeles, United States

A R T I C L E I N F O

Keywords:
Statistical learning
Transitional probability
Chunking
Embedded items
Illusory items
Infants

A B S T R A C T

Much research has documented infants' sensitivity to statistical regularities in auditory and visual inputs, however the manner in which infants process and represent statistically defined information remains unclear. Two types of models have been proposed to account for this sensitivity: statistical models, which posit that learners represent statistical relations between elements in the input; and chunking models, which posit that learners represent statistically-coherent units of information from the input. Here, we evaluated the fit of these two types of models to behavioral data that we obtained from 8-month-old infants across four visual sequence-learning experiments. Experiments examined infants' representations of two types of structures about which statistical and chunking models make contrasting predictions: illusory sequences (Experiment 1) and embedded sequences (Experiments 2–4). In all four experiments, infants discriminated between high probability sequences and low probability part-sequences, providing strong evidence of learning. Critically, infants also discriminated between high probability sequences and statistically-matched sequences (illusory sequences in Experiment 1, embedded sequences in Experiments 2–3), suggesting that infants learned coherent chunks of elements. Experiment 4 examined the temporal nature of chunking, and demonstrated that the fate of embedded chunks depends on amount of exposure. These studies contribute important new data on infants' visual statistical learning ability, and suggest that the representations that result from infants' visual statistical learning are best captured by chunking models.

1. Introduction

How do learners make sense of their intricately structured auditory and visual environments? Previous research suggests that both infants and adults are able to identify statistically coherent pieces of information contained within larger sequences presented both aurally and visually (see Krogh, Vlach, and Johnson (2012) for a review). This "statistical learning" ability may facilitate learners' detection of important types of environmental structure. For instance, statistical learning is thought to help learners identify words in continuous speech, facilitating language learning (e.g., Saffran, 2001), and help learners segment continuous motion into discrete events, facilitating visual learning and categorization (e.g., Stahl, Romberg, Roseberry, Golinkoff, & Hirsh-Pasek, 2014).

Despite the scope and potential importance of statistical learning ability, the specific processes underlying statistical learning remain unclear. Recent research has investigated how two types of models of the mechanisms underlying statistical learning – statistical models and chunking models – account for adults' statistical learning performances (see Thiessen, Kronstein, and Hufnagle (2013) for a review). Though several studies suggested that adults' statistical learning is best accounted for by chunking models (Fiser & Aslin, 2005; Giroux & Rey, 2009; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Poulin-Charronnat, 2012; Slone & Johnson, 2015b), at least one study suggested that statistical models may in some cases provide a better fit for adult data (Endress & Mehler, 2009). Moreover, it remains unknown which type of model best accounts for *infants'* statistical learning performances. This is an important issue to address, as statistical learning is posited to underlie much early learning, including language acquisition. Additionally, examining possible statistical and chunking processes in infants' statistical learning allows investigation of the extent to which the mechanisms underlying statistical learning are similar for infants and adults.

1.1. Statistical and chunking models of statistical learning

In a seminal study of infants' statistical learning, Saffran, Aslin, and Newport (1996) exposed 8-month-olds to a continuous stream of speech

* Corresponding author at: Indiana University Bloomington, Department of Psychological and Brain Sciences, 1101 E. Tenth Street, Bloomington, IN 47405-7007, USA.
E-mail address: laurenkslone@gmail.com (L.K. Slone).

https://doi.org/10.1016/j.cognition.2018.05.016
Received 22 June 2015; Received in revised form 18 May 2018; Accepted 21 May 2018

in an artificial language composed of four, three-syllable nonsense "words." Words were concatenated with no pauses between words, such that word boundaries were marked only by differing statistical relations between syllables within words and between words. After only 2 min of exposure to this language, infants were able to distinguish words from "part-words" (syllable sequences spanning word boundaries), demonstrating sensitivity to the statistical structure of the input.

Both statistical (or "transition-finding") models and chunking (or "clustering") models (Thiessen et al., 2013) have been proposed to account for such sensitivity to statistical structure (Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Thiessen et al., 2013); however, these models differ in the representations stored in memory. Statistical models can be instantiated with simple recurrent networks (SRNs) (e.g., Elman, 1990) that calculate and represent in memory statistical relations between items. For instance, one statistical relation that models (and human learners) may represent is transitional probability (TP), defined as the probability of event Y given event X ($P(Y|X)$), a measure of the strength with which X predicts Y. Representing such a statistic would not only inform the model of the likelihood of two items occurring together, but would also allow the model to predict individual items based on previous items in a sequence.

Consider, for instance, the Saffran et al. (1996) sequence composed of four 3-syllable words: $A_1A_2A_3$, $B_1B_2B_3$, $C_1C_2C_3$, and $D_1D_2D_3$. Statistical models will learn that $P(A_2|A_1)$ and $P(A_3|A_2)$ are high because items A_1, A_2, and A_3 always appear together in that order. In contrast, $P(B_1|A_3)$ will be lower because word A is sometimes followed by word B, but other times followed by words C or D. In this way, statistical models can distinguish statistically coherent units of information contained within a sequence (e.g., $A_1A_2A_3$) from less coherent units like part-words (e.g., $A_3B_1B_2$). Statistical models do not explicitly represent statistically coherent units; rather, they represent statistical relations between items.

In contrast, chunking models typically consider sensitivity to statistical relations like TPs to simply be a byproduct of other processes. The general feature separating chunking from statistical models is that chunking models *do* represent statistically coherent units of information in memory. The mechanisms by which chunking models acquire these representations differ across models. Some of the most common chunking models are Bayesian models (e.g., Goldwater, Griffiths, & Johnson, 2006, 2009; Orbán et al., 2008) and PARSER (Perruchet & Vinter, 1998). Despite their varying learning processes, the representations that result from chunking models are discrete, statistically coherent, "chunks" of information.

PARSER (Perruchet & Vinter, 1998), for instance, is a chunking model designed to account for human behavior by implementing psychologically plausible processes of attention, memory, and associative learning. PARSER joins items perceived within one attentional focus into a representational unit, or chunk. Representations of units whose component items co-occur regularly are progressively strengthened in memory, while representations of units whose component items do not co-occur regularly are forgotten. For instance, consider again the Saffran et al. (1996) sequence, and suppose that at any moment PARSER can only capture up to two items in its attentional focus. PARSER might initially capture the sequence $A_1A_2A_3B_1B_2B_3$ in three separate chunks: A_1A_2, A_3B_1, and B_2B_3. Over time, chunks A_1A_2 and B_2B_3 will be reinforced in memory because their component items always co-occur. In contrast, chunk A_3B_1 will only be weakly represented because its component items co-occur less frequently. Moreover, once the sequence A_1A_2 is represented as a single chunk rather than as two separate items, it becomes possible for the structure $A_1A_2A_3$ to be captured in a single attentional focus (i.e., as the aggregate of two items: A_1A_2 and A_3). Thus, with sufficient exposure PARSER will form strong representations of statistically coherent units of information (e.g., $A_1A_2A_3$) and distinguish them from weakly represented part-words (e.g., $A_3B_1B_2$).

1.2. Examining model fit to human data

Recent research has investigated how well statistical and chunking models fit human data (e.g., Endress & Mehler, 2009; Fiser & Aslin, 2005; Frank et al., 2010; Giroux & Rey, 2009; Orbán et al., 2008; Perruchet & Poulin-Charronnat, 2012; Slone & Johnson, 2015b). Many of these studies have examined the representations that adults store following auditory or visual statistical learning tasks, and whether these representations are best captured by statistical or chunking models. Such studies investigate representations of two types of items. The first type is illusory (or "phantom") units—units that are never presented to participants, but have the same statistical structure as other units that are presented. For example, if *tazepi*, *mizeru*, and *tanoru* are words presented in a speech stream, and TPs are .50 between syllables within these words (e.g., between *ta* and *ze* and between *ze* and *ru*), a statistically matched illusory word would be *tazeru* (Endress & Mehler, 2009). Statistical models could learn, for instance, that $P(ze|ta) = P(ru|ze) = P(pi|ze)$, and would therefore predict that the unit *tazepi* should be indistinguishable from the illusory word *tazeru* because the two strings are statistically equivalent. Chunking models, in contrast, predict that learners should fail to recognize illusory units because these units have never been presented, therefore learners could not have extracted from the input a chunk matching an illusory unit.

The second type of item researchers have investigated is embedded units—sub-units that occur only within larger units (Fiser & Aslin, 2005). In terms of linguistic materials, an embedded item could be a group of syllables that occurs within a word, but never occurs independently (e.g., "eleph", as in "elephant") (Thiessen et al., 2013). Statistical models predict that, because learners represent statistical relations between all pairs of adjacent elements, as learners become familiar with a unit (e.g., $A_1A_2A_3$), distinguishing components embedded in that unit (e.g., A_1A_2) should improve relative to less statistically coherent configurations of elements (e.g., A_3B_1). Many chunking models, in contrast, predict that as learners become familiar with a unit, they should become *less* able to distinguish components embedded in that unit from less statistically coherent configurations of elements (see Giroux & Rey, 2009). This is because an assumption of many chunking models is economy of representation, instantiated as competition between chunks within memory (Fiser & Aslin, 2005; Orbán et al., 2008; Thiessen et al., 2013). For instance, as PARSER learns the unit structure $A_1A_2A_3$, not only will the chunk $A_1A_2A_3$ be strengthened in memory, but it will also interfere with memory for embedded chunk A_1A_2, progressively reducing accessibility to A_1A_2 (Giroux & Rey, 2009).

Six studies have recently investigated adults' representations of illusory and embedded units, and the ability of various models to account for this performance. Specifically, these studies have investigated adults' representations of illusory units presented in auditory sequences (Endress & Mehler, 2009; Perruchet & Poulin-Charronnat, 2012) and visual sequences (Slone & Johnson, 2015b), and embedded units presented in auditory sequences (Giroux & Rey, 2009), visual sequences (Slone & Johnson, 2015b), and visual scenes (Fiser & Aslin, 2005; Orbán et al., 2008). Though one study (Endress & Mehler, 2009) suggested that statistical models may provide a better fit for adult statistical learning performances, the majority of these studies suggest that adults' statistical learning is best accounted for by chunking models.

It remains unknown, however, which type of model best accounts for infants' statistical learning performances. Two major types of chunking models (Bayesian models and PARSER) rely on assumptions about learners' priors (e.g. Goldwater et al., 2006, 2009) and attention, memory, and associative learning (Perruchet & Vinter, 1998)—factors that likely change between infancy and adulthood. For instance, PARSER is typically endowed with the ability to process up to three chunks simultaneously. This parameter seems plausible for modeling adults' learning, as much research suggests that adult short-term and working memory capacities fall in the range of 3–5 chunks (see reviews by Cowan, 2001, 2010). However, recent research suggests that during