Statistical learning and memory

Ansgar D. Endress
Department of Psychology, City, University of London, United Kingdom
Lauren K. Slone
Department of Psychological and Brain Sciences, Indiana University, Bloomington, United States
Scott P. Johnson
Department of Psychology, University of California, Los Angeles, United States

Draft of June 1, 2020

Abstract

Learners often need to identify and remember recurring units in continuous sequences, but the underlying mechanisms are debated. A particularly prominent candidate mechanism relies on distributional statistics such as Transitional Probabilities (TPs). However, it is unclear what the outputs of statistical segmentation mechanisms are, and if learners store these outputs as discrete chunks in memory. We critically review the evidence for the possibility that statistically coherent items are stored in memory and outline difficulties in interpreting past research. We use Slone and Johnson's (2018) experiments as a case study to show that it is difficult to delineate the different mechanisms learners might use to solve a learning problem. Slone and Johnson (2018) reported that 8-month-old infants learned coherent chunks of shapes in visual sequences. Here, we describe an alternate interpretation of their findings based on a multiple-cue integration perspective. First, when multiple cues to statistical structure were available, infants' looking behavior seemed to track with the strength of the strongest one — backward TPs, suggesting that infants process multiple cues simultaneously and select the strongest one. Second, like adults, infants are exquisitely sensitive to chunks, but may require multiple cues to extract them. In Slone and Johnson's (2018) experiments, these cues were provided by immediate chunk repetitions during familiarization. Accordingly, infants showed strongest evidence of chunking following familiarization sequences in which immediate repetitions were more frequent. These interpretations provide a strong argument for infants' processing of multiple cues and the potential importance of multiple cues for chunk recognition in infancy.

Keywords: Statistical learning; language acquisition; word segmentation; serial memory

In many domains, we need to recognize units in continuous sequences. Among many other examples, speech is a continuous signal, but to understand any sentence, we need to recognize individual words within a continuous speech sequence; people move continuously through space, but to make sense of a person's behavior, we need to recognize individual (goal-directed) actions within the flow of motion (e.g., Newtson,

This research was support by NIH grant R01-HD073535 to SPJ.

1973; Newtson, Engquist, & Bois, 1977; Zacks & Tversky, 2001; Zacks & Swallow, 2007); musical pieces such as symphonies or operas can be annoyingly lengthy, but we readily recognize recurring melodic elements. While the problem of recognizing recurring units seems daunting enough, it is exacerbated for learners who have to *identify* the underlying units (e.g., words, actions or motifs) to begin with. This is called the segmentation problem: Learners have to identify where unknown recurring units start and end in a continuous signal, and have then to commit them to memory.

While it is uncontroversial that humans and other animals learn *something* from such continuous sequences, it is debated whether what they learn is amenable to being placed in memory as a discrete chunk of material (e.g., a word or object). (In the following, "chunks" refer to units that are placed in memory; we refer to "segmentation" as the set of processes that individuate units and place them in memory.) The arguably most prominent proposal for how the segmentation problem might be solved relies on statistical learning.

Here, we critically review recent data that bears on the issue of whether statistical learning allows learners to place chunks in memory, and outline some difficulties in interpreting extant data, in particular due to the multitude of powerful learning mechanisms that infants have at their disposal and that they might use for different aspects of a learning problem. We conclude that statistical learning might be part of a suite of cognitive mechanisms for chunking items into units in continuous sequences, but that its contribution still needs to be delineated from that of other mechanisms that are also available to learners.

Statistical segmentation mechanisms

The arguably most prominent proposal for how the segmentation problem might be solved relies on Transitional Probabilities (TPs) among sequence elements. TPs are the conditional probabilities of the next element given the preceding element. For example, after hearing the syllable *whis*, we can predict the next syllable with higher certainty than the syllable following *key*, because TPs are thought to be higher within units (such as words like *whiskey*) than across unit boundaries. Human adults, infants, and many non-human animals are sensitive to TPs in a variety of domains (e.g., Aslin, Saffran, & Newport, 1998; J. Chen & Ten Cate, 2015; Creel, Newport, & Aslin, 2004; Endress, 2010; Endress & Wood, 2011; Fiser & Aslin, 2002, 2005; Glicksohn & Cohen, 2011; Hauser, Newport, & Aslin, 2001; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran & Griepentrog, 2001; Sohail & Johnson, 2016; Toro & Trobalón, 2005; Turk-Browne, Jungé, & Scholl, 2005; Turk-Browne & Scholl, 2009). As a result, learners might well use TPs to segment recurrent units from continuous sequences, a view that is also supported by a variety of computational models that are, in some form or another, based on similar distributional statistics (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Vinter, 1998; Swingley, 2005).

Prediction vs. memory

While the distributional statistics of a sequence tell learners which sequence elements are predictive of which other sequence elements, such knowledge does not imply that mutually predictable elements are memorized together. For example, we might well be able to predict the next syllable or word from the previous one, but this does not imply that the predictable syllable or word combinations are segmented and stored as single chunks in memory. After all, we can have a reasonable expectation that the syllable after *is* is the article *a* (which occurs after 12% of the occurrence of *is*; Davies, 2018), but this does not mean that *isa* is segmented as a word-like unit that is stored in memory. More generally, the dissociation between the ability to *predict* events and the ability to *store* event combinations in (declarative) memory is one of the classic findings in the neuropsychology of procedural memory (e.g., Knowlton & Squire, 1994; Knowlton, Mangels, & Squire, 1996; Poldrack et al., 2001). As a result, it is an empirical question whether (1) highly predictable sequences of syllables or other items are stored in memory (e.g., in the mental lexicon), and (2) whether the mechanisms that learn to predict items from one another are necessarily those that store the predictable sequences of items in memory. That is, even when it leads to successful prediction of items, statistical learning might not consistently yield the extraction and learning of genuine chunks that are present in the input and can be placed in memory.

Do statistical segmentation mechanisms lead to memories for chunks? Theoretical arguments

Before reviewing the empirical evidence on the question of whether distributional learning mechanisms lead to memory for chunks, it is useful to outline some theoretical desiderata for a mechanism that converts continuous sequences into memory representations for discrete chunks, using speech segmentation as a case study. Specifically, to the extent that the function of a segmentation mechanism is to learn words and to place them into what will be the mental lexicon, it needs to store recurrent syllable chunks in memory so that they can be retrieved when we try to produce or understand a word (Endress & Mehler, 2009b; Endress & Hauser, 2010; Endress & Langus, 2017). However, it is unclear whether TP-based learning mechanisms really accomplish this function, for three theoretical reasons.

First, and as mentioned above, the ability to store event combinations in (declarative) memory is dissociable from the ability to predict events from one another (e.g., Knowlton & Squire, 1994; Knowlton et al., 1996; Poldrack et al., 2001).

The second reason relates to how segmentation is usually assessed experimentally. Participants need to show discrimination between high-TP units and low-TP units, but, as mentioned above, such discriminations do not establish that the high-TP units have been learned as genuine chunks and are thus placed in memory. For example, adult and infant learners are sensitive to backward TPs (i.e., the conditional probability of an item given the *following* item; e.g. Hay, Pelucchi, Graf Estes, & Saffran, 2011; Perruchet & Desaulty, 2008; Pelucchi, Hay, & Saffran, 2009), and adult learners are just as good at discriminating high-TP units from low-TP units when these units are played forward as when they are played backwards (Endress & Wood, 2011; Turk-Browne & Scholl, 2009). However, while backward TPs provide better segmentation cues than forward TPs in some languages (e.g., Gervain & Guevara Erra, 2012; Saksida, Langus, & Nespor, 2017), learners clearly cannot place backward units in memory as they have never encountered them. Mirroring the dissociations between prediction and memory (e.g., Knowlton & Squire, 1994; Knowlton et al., 1996; Poldrack et al., 2001), successful discrimination between high-TP and low-TP units thus does not necessarily imply that the high-TP units have been learned as discrete chunks of material that will, eventually, populate a memory store such as the mental lexicon.

The same conclusion follows when considering possible mechanisms by which TPs might be tracked. For example, Endress and Johnson (under review) described a neural network where representations of syllables form associations (through Hebbian learning) if they are active simultaneously, and thus if they occur closely together in time. This model recognized high-TP items better than low-TP items, irrespective of whether the items were played forward or backward, without having a memory representation of either type of item.

The third reason for which it is unclear whether TP-based learning mechanisms lead to memorized chunks relates to the memory format of actual linguistic sequences.

It is well known that sequences can be encoded in different ways, notably by encoding the transitions between items or by encoding the positions of the different items with respect to items in the first and the last positions (i.e., the edges; see e.g., Fischer-Baum, Charny, & McCloskey, 2011; Henson, 1998, for reviews). While these two coding schemes are available to human and non-human learners (e.g., S. Chen, Swartz, & Terrace, 1997; J. Chen & Ten Cate, 2015; J. Chen, Jansen, & Ten Cate, 2016; Coye, Ouattara, Zuberbühler, & Lemasson, 2015; Endress, Carden, Versace, & Hauser, 2010; Marchetto & Bonatti, 2013, 2015; Seidl & Johnson, 2006, 2008; Sohail & Johnson, 2016), they seem to be independent of one another and show multiple dissociations, including their sensitivity to temporal order, the kinds of cues they require, the speed with which they can be learned from fluent speech, their developmental time course, and, in vision, their tolerance of viewpoint changes (Endress & Mehler, 2009a; Endress & Wood, 2011; Marchetto & Bonatti, 2013, 2015; Peña, Bonatti, Nespor, & Mehler, 2002; see Endress & Bonatti, 2016, for a review).

Critically, based on evidence from speech errors, reading, and deficits in brain damaged patients, it has been argued that linguistic sequences are encoded using the edge-based coding scheme rather than the transition-based coding scheme (e.g., Fischer-Baum, McCloskey, & Rapp, 2010; Fischer-Baum et al., 2011; Miozzo, Petrova, Fischer-Baum, & Peressotti, 2016). As a result, learners might need to rely on other cues to learn recurring units to make them compatible with those mechanisms that will store them in memory. In the case of speech segmentation, this is not necessarily a problem, as there are a number of other cues that are available across languages. These include mechanisms that use known words as anchors for segmentation (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Brent & Siskind, 2001; Mersad & Nazzi, 2012; see also Van de Weijer, 1999, but see Aslin, Woodward, LaMendola, & Bever, 1996), mechanisms that extract items from the beginnings and endings of utterances (e.g., Seidl & Johnson, 2006, 2008; Shukla, Nespor, & Mehler, 2007; Sohail & Johnson, 2016) and mechanisms that rely on universal aspects of prosody (e.g., Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Fenlon, Denmark, Campbell, & Woll, 2008; Pilon, 1981; Selkirk, 1984, 1986; for an overview see e.g., Cutler, Oahan, & van Donselaar, 1997; Langus, Marchetto, Bion, & Nespor, 2012; Shattuck-Hufnagel & Turk, 1996). Further, at least by 8 months, infants seem to be more sensitive to these speech cues than to statistical cues (e.g., Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Shukla, White, & Aslin, 2011).

For these reasons, one cannot take for granted that a sensitivity to TPs necessarily translates to creation of memory representations for the high-TP items, and it is critical to assess this question empirically.

Do statistical segmentation mechanisms lead to memories for chunks?

Empirical arguments

There are three literatures that seem to lead to contradictory conclusions about whether sequences that can be segmented based on TPs are actually encoded as a set of segmented chunks. One literature investigates the fate of high-TP items: Do high

TP-items have a subsequent processing advantage, for example for word-learning and memory? The second literature investigates the relative weight of the frequency of chunks compared to the TPs among the elements within these chunks. The third literature investigates to what extent learners are sensitive to sub-chunks once they are presumed to have learned chunks.[1] We will now discuss these literatures in turn.

Subsequent processing of high-TP items: The case of word learning

One way to address the question of whether high-TP items are memorized as chunks is to test whether it is easier to assign meanings to high-TP items compared to low-TP items. This line of research was initiated by Graf-Estes, Evans, Alibali, and Saffran's (2007) experiments. In their studies (see also Estes, 2012; Hay et al., 2011), infants were familiarized with a continuous speech stream comprised of nonsense words; as a result, TPs among syllables within these words were higher than TPs between syllables that straddled word boundaries. In a subsequent word-learning phase, infants had to associate visual images with words (i.e., high-TP units), non-words, or part-words (i.e., low-TP units). Learning of associations between visual images and sounds was better for high-TP units than for low-TP units.

At first sight, these results seem to suggest that the high-TP units were learned

[1] In the following section, we consider the frequency of chunks as a potential determinant of how well they are learned, but many other factors influence what chunks can be learned, including when they were last encountered (e.g., Bjork & Allen, 1970; Ebbinghaus, 1885/1913; Vlach & Johnson, 2013) and the extent to which they are similar to and interfere with other items (e.g., Berman, Jonides, & Lewis, 2009; Endress & Szabó, 2017), which in the case of lexical acquisition might be related to neighborhood density (e.g., Storkel & Lee, 2011). Here, we focus on chunk frequency because this is the cue that has been manipulated in the experiments reviewed below.

and memorized as chunks. However, there are several alternative interpretations that relate to the predictiveness of items. For example, if infants form associations between individual syllables and visual images, learning might be better for high-TP units, not because the high-TP units are represented as chunks, but rather because the second order associations between individual syllables and visual items are stronger in these units, so that syllables and visual items are more predictive of one another without the syllables being memorized as a coherent chunk (see Endress & Langus, 2017, for discussion and further alternative interpretations).

Karaman and Hay (2018) followed a different approach to show that a preference for high-TP items over low-TP items suggests that the high-TP items are retained in memory. They asked if high-TP items are preferentially consolidated in long-term memory compared to low-TP items. They first established that, after being familiarized with passages containing both high-TP and low-TP items, 8-month-olds do not discriminate high-TP items from low-TP items after a delay of just 10 min (though they do not discriminate these items immediately after the familiarization either, except in Experiment 3, where a different attention grabber was used). Critically, when infants were first familiarized with the passage, then tested immediately afterwards on high-TP items and low-TP items and then tested again 10 min later, discrimination is reestablished in the second, delayed test phase (though the preference for high-TP items differed neither from that on the immediate test, nor from that on the successful immediate test in Experiment 3).

Karaman and Hay's (2018) explanation for the successful discrimination in the second test proceeds in three steps. First, infants learn the TP structure of the items during the initial familiarization. Second, infants are exposed to both high-TP items and low-TP items during the initial test. Third, during the retention interval, infants preferentially consolidate high-TP items into long-term memory.

However, leaving aside the issue of whether these results are statistically reliable, a preference for high-TP items over low-TP items does not necessarily imply that the high-TP items are stored in long-term memory: French-learning infants prefer non-words whose syllables have high-TPs in French compared to non-words whose syllables have low-TPs in French (Ngon et al., 2013); unless infants have memory representations of these items they have never heard before, this seems to suggest that a discrimination between high-TP items and low-TP items does not imply memory for the high-TP items even when the discrimination is supported by some form of long-term memory.

Subsequent processing of high-TP items: Beyond word learning

Experiments inspired by Graf-Estes et al. (2007) have the drawback that the learning advantage for high-TP items might be due to second-order TPs. Several authors attempted to provide more direct evidence that high-TP items are memorized in Graf-Estes et al.'s (2007) paradigm.

For example, Erickson, Thiessen, and Estes (2014) tested the influence of the statistical properties of verbal labels in Waxman and Markow's (1995) category learning task. In that task, infants view some exemplars of a category (e.g., some dinosaurs) accompanied by a common verbal label; following this, they view a novel exemplar of the category and an exemplar of another category (e.g., a fish), and their preference is measured. For superordinate categories (e.g., animals, vehicles), infants require verbal labels to exhibit a novelty preference for the novel category; in contrast, for basic-level categories (e.g., cars, dogs), infants show a novelty preference also in the absence of verbal labels.

In Erickson et al.'s (2014) experiments, the labels were either high-TP items or low-TP items, taken from a speech stream with which infants were previously familiarized. Infants preferred the exemplar from the novel category when the label was a high-TP item but not when the label was a low-TP item. Critically, when infants were *not* pre-familiarized with the speech stream, they had no such novelty preference, suggesting that infants could use only high-TP items but not low-TP items as category labels (though these results contrast with Waxman and Markow's (1995), where infants showed a novelty preference for basic level categories even in the absence of verbal labels). If so, a plausible conclusion is that the high-TP items have been memorized.

However, this conclusion is not necessarily warranted, for two reasons. First, infants listened longer to the high-TP items when they were presented together with the category exemplars during training; hence, any preference for the novel category with the high-TP items might reflect the greater exposure to the old category when it was accompanied by high-TP items. Second, Erickson et al. (2014) would predict an interaction between label type (high-TP vs. low-TP item) and familiarization type (familiarization vs. no familiarization); while they do not report this interaction, it is unlikely to be significant.[2]

Critically, leaving aside the issue of whether these results are reliable, the second-order associations between individual syllables and category representations might be stronger for high-TP items, just as the second-order associations between syllables and visual items might be stronger for high-TP items in Graf-Estes et al.'s (2007) experiments, suggesting that Erickson et al.'s (2014) experiment are open to similar alternative interpretations as Graf-Estes et al.'s (2007).

Shoaib, Wang, Hay, and Lany (2018) leveraged individual learning differences to provide evidence that high-TP items are stored in memory. They predicted that infants with greater vocabularies should be worse at learning words with non-native phonotactics than infants with smaller vocabularies. (It is not entirely clear what motivates this hypothesis: after all, phonotactic knowledge is acquired very early in life when lexical knowledge is presumably limited (e.g., Jusczyk, 1999); further, it is not clear how phonotactic knowledge might be used in the service of word learning if it requires lexical knowledge to begin with.) In an experiment similar to Graf-Estes et al.'s (2007), Shoaib et al. (2018) familiarized 20-month-old English learning infants with one of two Italian passages containing high-TP words and low-TP words.

[2] Simulations with 10,000 Gaussian random samples based on Erickson et al.'s (2014) means and standard deviations suggest that this interaction is unlikely to be significant (mean $F(1,50) = 2.76$, $p = .103$; median $F(1,50) = 1.08$, $p = .303$).

Following this familiarization, infants were presented with pictures of unfamiliar animals accompanied by a high-TP label or a low-TP label. Finally, during test, infants viewed two animals presented side by side, accompanied by a label. Shoaib et al. (2018) asked whether infants would look longer at the labeled animal.

Results revealed that infants preferentially looked at the labeled animal for both high-TP labels and low-TP labels, but only when familiarized with one of the passages; infants familiarized with the other passage showed no evidence of learning. Critically, for high-TP labels, there was a negative correlation between vocabulary size and preference for the labeled animal. However, inspection of their Figure 3 suggests that this correlation is mostly driven by outliers. When infants who differ more than 2.5 standard deviations from the mean are removed (either in Accuracy, $N = 1$, or in Vocabulary, $N = 2$, based on data digitized using `https://apps.automeris.io/wpd/`), the initially highly significant correlation ($p = .0009$) is no longer significant ($p = .095$).

Leaving aside questions of data reliability, there are two mutually non-exclusive alternative interpretations of Shoaib et al.'s (2018) results. First, infants with larger vocabularies might be faster learners. As a result, they might quickly identify the labeled animal, and then start exploring the screen. As looks are measured in a 1.5 s interval starting 300 ms after the onset of the label, this would lead high vocabulary infants to appear to have a relatively low propensity to look at the matching animals. Second, while infants were certainly unlikely to know labels for animals such as armadillos, these animals might look similar enough to other animals for which high vocabulary infants are more likely to know labels; as a result, the mutual exclusivity principle (e.g., Markman, 1994; Halberda, 2003) would make it harder for these infants to acquire a second label on top of the one they already know. As a result, the question of whether high-TP items are stored in memory is still open.

The relative weight of TPs and frequency

Another strategy to test if learners store the output of statistical segmentation computations as chunks is to assess the relative importance of TPs and chunk frequency. If learners memorize chunks, they should be more familiar with chunks they have memorized than with items they have not encountered, even if they are favored by TPs.

Endress and Mehler (2009b) followed this strategy by exposing adult participants to a speech stream consisting of a random-arrangement of six syllable triplets (hereafter called "words"). As in other statistical learning experiments, TPs within words were higher than TPs across word boundaries. Critically, these words were constructed so that there were "illusory words" that had exactly the same TPs as the words that appeared in the familiarization stream, but were never actually encountered (i.e., the illusory words had a frequency of 0).

Following this familiarization, participants had to choose between words (high-TP syllable triplets that had occurred in the speech stream), illusory words (high-TP syllable triplets that had not occurred in the speech stream), and part-words (low-

TP syllable triplets that had occurred in the speech stream but straddled a word boundary and occurred less frequently than words).

Endress and Mehler (2009b) found three crucial results (see Endress & Langus, 2017, for a replication in the visual modality). First, participants considered illusory words as more familiar than part-words, suggesting that they were more sensitive to the higher TPs of illusory words than to the higher frequency of part-words. Second, their preference for words over part-words was much more pronounced than that for words over illusory words, suggesting again that even relatively subtle differences in TPs count more than massive differences in unit frequency. Third, at least in Endress and Mehler's (2009b) study (but see below), participants had no preference for words over illusory words at all, which would suggest that they had little sensitivity to chunk frequency.

In a similar study, Perruchet and Poulin-Charronnat (2012) also found that participants are more sensitive to TPs than to the frequency of chunks, but, in their experiments, participants preferred words over illusory words, suggesting that they had some sensitivity to chunk frequency.

These findings are consistent with a model like that of Endress and Johnson's (under review) model, which predicts the greater importance of TPs compared to chunk frequency. Because it is based on pairwise associations among syllables, it prefers illusory words to part-words, but does not discriminate between actual words and illusory words (since it does not represent chunks in memory).

The role of additional cues

These results do not imply that learners are never sensitive to chunk frequency; rather they might simply need different types of cues. In fact, infants excel at exploiting multiple cues when they are available (e.g., Frank, Slemmer, Marcus, & Johnson, 2009; Gerken, Wilson, & Lewis, 2005; Schonberg, Marcus, & Johnson, 2018; Ter Schure, Mandell, Escudero, Raijmakers, & Johnson, 2014). Accordingly, both Endress and Mehler (2009b) and Endress and Langus (2017) showed that adult participants prefer words to illusory words when additional cues are given, such as prosodic cues indicating the beginning and the end of each word. Endress and Langus (2017) suggested that such cues established a sensitivity to chunk frequency because it enabled the kind of edge-based encoding thought to underlie the representation of linguistic sequences (e.g., Fischer-Baum et al., 2010, 2011; Fischer-Baum & McCloskey, 2015).

This conclusion is also in line with other experiments in the visual modality. For example, when using visual stimuli where each element was a shape/location combination rather than a syllable, Slone and Johnson (2015) reported a preference for the visual equivalent of words over the visual equivalent of illusory words. In this case, an extra cue might have been provided by the spatial trajectory inherent in each unit. For example, if the visual sequence included a "word" (a high-TP three-shape/location combination) with the shape/location combinations *ABC*, and if *A* appeared in the upper left corner, *B* in the middle, and *C* in the upper right corner, the word *ABC* would generate a V-shaped trajectory. These trajectories might have allowed participants to discriminate words from illusory words, as these trajectories were only experienced for words (though TPs in words and illusory words were identical). If so, these results would corroborate the possibility that additional cues can help learners establish a greater sensitivity to chunk frequency. Importantly, however, the evidence overall suggests that, when cues in addition to TPs are unavailable, learners are more sensitive to TPs than to chunk frequency, which is highly problematic for any model that assumes that the output of TPs is learned chunks.

Do learners recognize sub-units of units in vision?

While the literature on illusory words suggests that sequences that can be segmented based *only* on TPs may not be stored in memory as chunks of material, the visual statistical learning literature seems to lead to the opposite conclusion. Specifically, a number of experiments suggest that, once (presumed) chunks or units are learned, sub-units become less accessible. To use a linguistic analogy, when hearing the word *hamster*, it is difficult to recognize that the first syllable is a word on its own (i.e., *ham*), though, in the case of word recognition, such effects are driven at least in part by phonetic differences between syllables that are parts of words and syllables that are words on their own (e.g., van Alphen & van Berkum, 2010; Salverda, Dahan, & McQueen, 2003; Shatzman & McQueen, 2006a, 2006b).

Similar effects have been observed in visual statistical learning of simultaneously presented shapes (Fiser & Aslin, 2005; Orbán et al., 2008): Entire units are easier to recognize than subunits. If so, the entire units are presumably stored in memory as chunks. Below, we will call this phenomenon the *sub-unit effect*.

However, it is unclear how reliable the sub-unit effect is. For example, Fiser and Aslin (2005) observed it in their Experiments 1 and 4, but not in their Experiment 5, and, when presenting shapes in a sequence rather than simultaneously, Slone and Johnson (2015) also failed to find evidence for the sub-unit effect in their Experiment 2, where they directly contrasted the strength of representation of units vs. sub-units. In fact, in unpublished results, we found that, at least with simultaneously presented shapes, the sub-unit effect can be replicated when the sub-units happen to be those parts of a unit that do not attract attention, but not when the sub-units come from salient parts of the units (Endress, in preparation). As a result, the sub-unit effect might be due to the perceptual organization of visual scenes rather than to statistical learning *per se*.

Chunks vs. TPs in infancy

The discussion about the relative weight of TPs and chunk frequency and the fate of sub-units above relied mostly on research with adults. We will now take advantage of Slone and Johnson's (2018) recent experiments addressing these issues in infancy to illustrate some of the interpretative difficulties outlined above.

Slone and Johnson (2018) asked if infants would discriminate visual words from illusory words (Experiment 1) and discriminate units from sub-units (Experiment 2 to 4). In all experiments, Slone and Johnson (2018) used sequentially presented stimuli, where each stimulus was a shape/location combination.

Slone and Johnson's (2018) Experiment 1 was inspired by Endress and Mehler's (2009b) finding that adults have much more difficulty discriminating units that differ only in frequency than discriminating units that differ only in TPs (Endress & Mehler, 2009b; Endress & Langus, 2017; Perruchet & Poulin-Charronnat, 2012), to the extent that, in some experiments but not others, listeners do not appear to be sensitive to unit frequency at all (see above).

Infants were exposed to a continuous sequence of 5 shapes where each shape appeared at a distinct spatial location, as in Slone and Johnson's (2015) adult study described previously. Critically, the shape/location combinations were arranged into three recurring units that played the role of words in speech segmentation experiments, and that were presented in a random order (see Figure 1). Two of these units were triplets (*ABC* and *DAC*, where each letter stands for a shape/location combination) and one was a pair (*BE*). As a result, the TPs among shape/location combinations were either 1.0 or .50 within units and .33 across units. Following this familiarization, infants were presented with three types of test sequences. One was a triplet encountered during the familiarization (*ABC*), one was an illusory triplet that did not appear during the familiarization but had the same TPs as the actual triplet (*ABE*), and one was a part-triplet (*CBE*), with weaker TPs than in the other two trial types.

Infants looked longest to the actual triplets, followed by the part-triplets, and

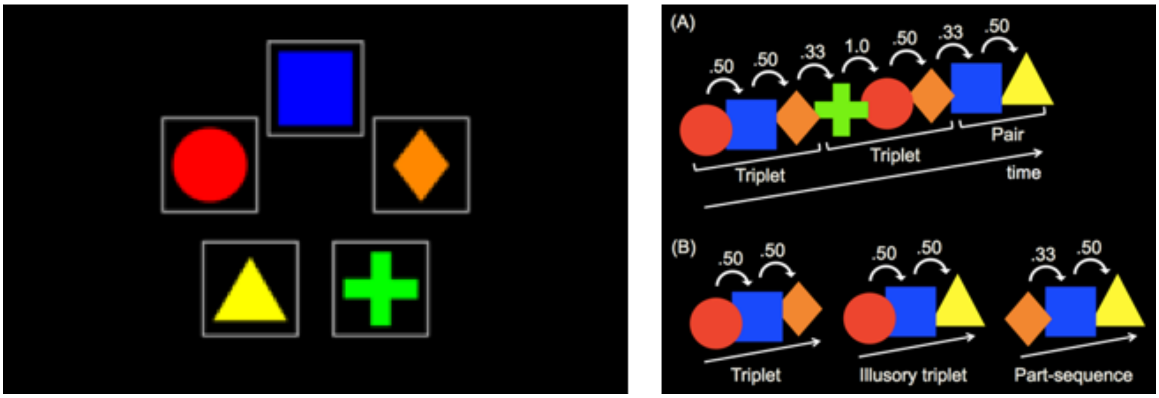

Figure 1. Left: Schematic depiction of the spatial array and shapes presented in Experiments 1–3 of Slone and Johnson (2018). Only one shape appeared at a time during familiarization and test. Right: Example (A) familiarization sequence and (B) test sequences presented in Experiment 1 of Slone and Johnson (2018). Numbers above shapes represent TPs during familiarization. Brackets below shapes indicate the unit structure of the familiarization sequence.

looked least to the illusory triplets. In contrast to earlier research where infants were unable to segment words from fluent speech when the words differed in length (Johnson & Tyler, 2010), infants thus looked longest to the sequences they were (presumably) most familiar with. Critically, they discriminated between actual and illusory triplets, suggesting that they might be sensitive to the chunk frequency of the sequences.

Slone and Johnson's (2018) Experiments 2 to 4 were inspired by earlier work showing that learners are more sensitive to entire units than to sub-units (Fiser & Aslin, 2005; Orbán et al., 2008; see above). In Experiments 2 and 3 (which were replications of one another), infants were again familiarized with a sequence of shape/location combinations. The shape/location combinations were arranged into a triplet (*ABC*) and a pair (*DE*) that were presented in random order for 80 times each.

Following this familiarization, infants were exposed to three types of test sequences: the high-TP pair from the familiarization sequence (*DE*), a low-TP part-pair that comprised two items crossing a unit boundary (*CD*), and a high-TP "embedded" pair that was a sub-unit of the triplet *ABC* (*BC*). Despite the actual pair and the embedded pair having identical TPs, infants' looking times were shorter for the actual pair compared to either the part pair or the embedded pair, suggesting that they were more familiar with the actual pair than with the other test sequences. Slone and Johnson (2018) concluded that, given that participants discriminated between pairs and embedded pairs, they must have extracted pairs (and triplets) as chunks.

Experiment 4 was similar to Experiments 2 and 3 except that two triplets and two pairs were used; in other words, the number of units to be learned was doubled, such that infants were exposed to each (triplet or pair) unit half as many times as in Experiments 2 and 3. Under these conditions, infants looked longer at part pairs than at either actual or embedded pairs, implying that the infants did learn the TP differences among adjacent shape/location combinations but did not learn the triplet chunks. Slone and Johnson (2018) hypothesized that the sub-unit effect did not emerge because, with this more complex material, infants did not have time to learn the chunk structure in the sequence.

Do infants represent statistical sequences as chunks? One interpretation of these results is that, in contrast to similar studies with adults, infants exposed to TP-based sequences extract and learn chunks: They discriminate actual items from illusory items, and they show the sub-unit effect.

However, a closer look at the results reveals that infants behaved in subtly different ways across experiments; for example, they exhibited different directions of preference across the different experiments. Given that even the highly simplified learning situations from word segmentation experiments recruit multiple distinct learning mechanisms (Endress & Bonatti, 2016; Peña et al., 2002), and that infants excel at exploiting multiple cues that are presented to them (e.g., Frank et al., 2009; Gerken et al., 2005; Schonberg et al., 2018; Ter Schure et al., 2014), it is interesting to consider the possibility that infants deploy different learning mechanisms in different situations.

Specifically, in Experiment 1, infants looked longer to familiar items at test, but in Experiment 2 to 4, infants looked longer to unfamiliar items. As noted by Slone and Johnson (2018), these different directions of preference may have stemmed from different degrees of habituation to the familiarization stimulus across experiments, as fewer infants habituated in Experiment 1 (11% of the infants) compared to Experiments 2, 3, and 4 (56%, 35%, and 44%, respectively). That is, infants in Experiment 1 might have been more liable to show familiarity preferences following the familiarization sequence, as few were habituated. However, Slone and Johnson (2018) found no correlation between the degree of habituation in individual infants and the degree of novelty preference, and the absence of a correlation was 7.6 more likely than its presence after correction with the Bayesian Information Criterion, 2.4

more likely after correction with the Akaike Information Criterion (Glover & Dixon, 2004).

The role of backwards TPs. An alternative explanation for the switch from a familiarity preference in Experiment 1 to a novelty preference in Experiments 2 through 4 is that infants track multiple cues, and that their behavior is driven by the *strongest* available cue. One such cue might be backward TPs. Learners are clearly sensitive to backward TPs (e.g., Hay et al., 2011; Perruchet & Desaulty, 2008; Pelucchi et al., 2009), and it turns out that backwards TPs were strongest in Experiment 1, but not in the other experiments. Specifically, and as shown in Figure 2, the backward TPs in illusory triplets were .5 and 1.0, those in part triplets .33 and 1.0, and those in actual triplets .5 and .5. Given that the differences in forward TPs were relatively subtle while the differences in backward TPs were relatively strong, infants might have adaptively chosen the strongest available cue, and learned the backward TPs. If so, they should be most familiar with illusory triplets, followed by part triplets and then by actual triplets, which seems to reflect the actual results, assuming that infants look longest to the items they are *least* familiar with. Infants might thus have a novelty preference in all four experiments, such that their behavior was driven by backwards TPs when this was the strongest available cue (i.e., in Experiment 1), and by other cues in Experiments 2 to 4.

This interpretation also explains why, in previous experiments using illusory triplets, the discrimination based on TPs was easier than discrimination based on

Forward items			**Backward items**		
Units					
Unit	**Possible continuations**	**Probability**	**Unit**	**Possible continuations**	**Probability**
ABC	→ ABC	33.3%	CBA	→ CBA	33.3%
	→ DAC	33.3%		→ CAD	33.3%
	→ BE	33.3%		→ EB	33.3%
DAC	→ ABC	33.3%	CAD	→ CBA	33.3%
	→ DAC	33.3%		→ CAD	33.3%
	→ BE	33.3%		→ EB	33.3%
BE	→ ABC	33.3%	EB	→ CBA	33.3%
	→ DAC	33.3%		→ CAD	33.3%
	→ BE	33.3%		→ EB	33.3%

Shape/Location Combination (SLC)						
SLC	**Possible continuations**	**Probability**	**SLC**	**Possible continuations**		**Probability**
A	→ B	50.0%	A	→ (after backward triplet CBA)	→ C	33.3%
	→ C	50.0%			→ E	16.7%
B	→ C	50.0%		→ (in backward triplet CAD)	D	50.0%
	→ E	50.0%	B	→ (in backward triplet CBA)	A	50.0%
C	→ A	33.3%		→ (after backward pair EB)	→ C	33.3%
	→ D	33.3%			→ E	16.7%
	→ B	33.3%	C	→ B		50.0%
D	→ A	100.0%		→ A		50.0%
E	→ A	33.3%	D	→ C		66.7%
	→ D	33.3%		→ E		33.3%
	→ B	33.3%	E	→ B		100.0%

Test items						
		TP (1st to 2nd SLC)	**TP (2nd to 3rd SLC)**		**TP (1st to 2nd SLC)**	**TP (2nd to 3rd SLC)**
Triplet	ABC	50.0%	50.0%	CBA	50.0%	50.0%
Illusory triplet	ABE	50.0%	50.0%	EBA	100.0%	50.0%
Part triplet	CBE	33.3%	50.0%	EBC	100.0%	33.3%

Figure 2. Forward and backward transitional probabilities in Slone and Johnson's (2018) Experiment 1. (Top) The top panel shows the units in the familiarization stream and their possible transitions, for forward items (left) and backward items (right). Each letter stands for a shape/location combination. (Middle). Possible transitions between individual shape/location combinations during familiarization, for forward transitions (left) and backward transitions (right). (Bottom). Transitional probabilities among shape/location combinations in the test items.

frequency (Endress & Mehler, 2009b; Endress & Langus, 2017; Perruchet & Poulin-Charronnat, 2012), while Slone and Johnson's (2018) Experiment 1 seems to suggest the opposite, as backward TPs were not informative in these earlier experiments.[3]

[3] A multiple cue integration perspective suggests another, not mutually exclusive, interpretation of Slone and Johnson's (2018) Experiment 1. As mentioned above, Slone and Johnson's (2015) (adult) participants might have preferred actual triplets to illusory triplets due to the unique spatial trajectory of triplets as opposed to illusory triplets; this view is particularly plausible because, at least in Working Memory, objects seem to get bound to their location in a fairly automatic way (e.g., Makovski & Jiang, 2008; Makovski, 2016). A similar explanation might thus hold for Slone and Johnson's (2018) Experiment 1 as well: Infants might discriminate actual triplets from illusory triplets because these items imply different spatial trajectories. Note that, as only the trajectory of the actual triplets would be familiar to infants, this explanation would support the interpretation that infants showed a familiarity preference in Experiment 1. In contrast, spatial trajectory is an unlikely explanation for infant preference in Experiments 2 to 4, as the trajectories implied by actual

Immediate repetitions of chunks. In Slone and Johnson's (2018) Experiments 2 and 3, infants discriminated between units and sub-units, providing evidence for the sub-unit effect and strongly suggesting that infants extracted and learned chunk-like units. The critical question is *how* they achieved this feat: Did they just use TP-based computations, or did they resort to other strategies?

While infants might have used TP-based computations, there is one feature of Slone and Johnson's (2018) experiments that may lie at the heart of the ability to chunk items: In their experiments, units occurred in immediate repetition, which might lead infants to extract units as chunks in turn. For example, when we see the sequence $ABCABCABC\ldots$, interrupted by some other sequence, we are likely able to memorize the chunk ABC, not because we compute the TPs between the A's, B's and C's (though we might do so as well), but rather because the immediate chunk repetition aids memorization. Such repetitions occurred with considerable probability in familiarization sequences such as Slone and Johnson's (2018) (Weisstein, n.d.).

This interpretation is consistent with earlier results. First, it is known since Ebbinghaus (1885/1913) that repeating items aids memorization. Second, in earlier speech segmentation experiments, the inclusion of immediate item repetitions enabled computations that were not available in the absence of such repetitions (Bonatti, Peña, Nespor, & Mehler, 2005). Third, there are both empirical and theoretical reasons to think that immediate repetitions of items might be particularly helpful for extracting chunks (Onnis, Waterfall, & Edelman, 2008; Vlach & Johnson, 2013).

pairs and embedded pairs should be equally familiar.

Further, once a chunk has been learned, it can be recognized in the sequence, which, in turn, likely improves segmentation performance for the items that have not been recognized yet (Bortfeld et al., 2005; see also Kim & Sundara, 2015; Lew-Williams, Pelucchi, & Saffran, 2011; Shi & Lepage, 2008).

This view may also help to explain why infants in Experiment 4 did not discriminate actual pairs from embedded pairs: given that Experiment 4 used more items, these items were less likely to occur in immediate repetition, and thus may have been less likely to be learned as *chunks* (see also Vlach & Johnson, 2013). In contrast, TP-based computations seem less sensitive to the presence or absence of immediately repeated items. As a result, infants successfully discriminated high-TP items from low-TP items, and looked longer to part-sequences (which had lower TPs) than to either true pairs or embedded pairs. In other words, infants are exquisitely sensitive to chunks when presented with continuous sequences. However, like adults, they might require specific cues such as the immediate repetition of items to extract the relevant chunks, while TP-based computations may operate irrespective of the presence of other cues (see Endress & Bonatti, 2016 for more results supporting this conclusion).

Conclusions

An important question in cognitive development is how we extract and learn chunks of recurring materials in continuous sequences. We critically reviewed evidence that one of the most prominent candidate mechanisms — statistical learning based on TPs — might serve this purpose. The evidence is decidedly mixed. While some results suggest that learners can use TPs to place chunks of recurring material in memory, we showed that these results have alternative interpretations and are not always statistically reliable.

Interpretation of such data is particularly difficult because learners can process multiple cues simultaneously (e.g., Gervain & Endress, 2017; Endress & Bonatti, 2016; Frank et al., 2009; Gerken et al., 2005; Schonberg et al., 2018; Ter Schure et al., 2014), and might rely on the most reliable (e.g., Frank & Tenenbaum, 2011; Gerken, 2010) or the most salient cue (Endress, 2013; Gervain & Endress, 2017). A case in point is Slone and Johnson's (2018) Experiment 1, where backward TPs are the strongest available cue, and the infants' looking behavior seemed to track the strength of backward TPs in the different test items rather than the frequency of recurring chunks. This does not necessarily imply that forward and backward TPs are tracked by separate mechanisms, and, in fact, a mechanism as simple as Hebbian correlational learning can handle TPs in both directions, (Endress & Johnson, under review). It does imply, however, that one must take care to test for and exclude alternative mechanisms, similarly to how cognition is studied in other animals (e.g., van Heijningen, de Visser, Zuidema, & ten Cate, 2009; Shettleworth, 2010).

Further, learners may *need* multiple cues to solve a learning problem (e.g., Gerken et al., 2005; Endress & Langus, 2017). Slone and Johnson's (2018) experiments provide a case in point for this possibility as well. In their Experiments 2 and

3, infants could rely on both TPs and immediate chunk repetitions during familiarization and showed evidence for extracting chunks. In contrast, in their Experiment 4, where chunk repetitions were less systematic, no evidence for chunking was obtained.

Taken together, this discussion raises two critical questions for future studies of statistical learning: which cues allow learners to memorize chunks from continuous sequences, and whether these cues are available in naturalistic input.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In K. Demuth & J. L. Morgan (Eds.), *Signal to syntax. bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167-206.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 317-33. doi: 10.1037/a0014873

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 567–572. doi: https://doi.org/10.1016/S0022-5371(70)80103-7

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(8), 451-459.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304. doi: 10.1111/j.0956-7976.2005.01531.x

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93-125.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-44.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72.

Chen, J., Jansen, N., & Ten Cate, C. (2016). Zebra finches are able to learn affixation-like patterns. *Animal Cognition*, *19*(1), 65–73. doi: 10.1007/s10071-015-0913-x

Chen, J., & Ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes*, *117*, 29–34. doi: 10.1016/j.beproc.2014.09.004

Chen, S., Swartz, K., & Terrace, H. S. (1997). Knowledge of the ordinal position of list items in rhesus monkeys. *Psychological Science*, *8*, 80–6.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268.

Coye, C., Ouattara, K., Zuberbühler, K., & Lemasson, A. (2015). Suffixation influences receivers' behaviour in non-human primates. *Proceedings. Biological Sciences*, *282*(1807). doi: 10.1098/rspb.2015.0265

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(5), 1119-30. doi: 10.1037/0278-7393.30.5.1119

Cutler, A., Oahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

Davies, M. (2018). *The 14 billion word iweb corpus.*

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University. (http://psychclassics.yorku.ca/Ebbinghaus/)

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–190.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176. doi: 10.1016/j.cognition.2012.11.014

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(1), 19–35.

Endress, A. D., Carden, S., Versace, E., & Hauser, M. D. (2010). The apes' edge: positional learning in chimpanzees and humans. *Animal Cognition*, *13*(3), 483-495. doi: 10.1007/s10071-009-0299-8

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177-199.

Endress, A. D., & Johnson, S. P. (under review). When forgetting fosters learning: A neural network model for statistical learning.

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. doi: 10.1016/j.cogpsych.2016.11.004

Endress, A. D., & Mehler, J. (2009a). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209.

Endress, A. D., & Mehler, J. (2009b). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351-367.

Endress, A. D., & Szabó, S. (2017). Interference and memory capacity limitations. *Psychological Review*, *124*(5), 551–571. doi: 10.1037/rev0000071

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171.

Erickson, L. C., Thiessen, E. D., & Estes, K. G. (2014). Statistically coherent labels facilitate categorization in 8-month-olds. *Journal of Memory and Language*, *72*, 49–58. doi: 10.1016/j.jml.2014.01.002

Estes, K. G. (2012). Infants generalize representations of statistically segmented words. *Frontiers in Psychology*, *3*. doi: 10.3389/fpsyg.2012.00447

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177-200.

Fischer-Baum, S., Charny, J., & McCloskey, M. (2011). Both-edges representation of letter position in reading. *Psychonomic Bulletin and Review*, *18*(6), 1083–1089. doi: 10.3758/s13423-011-0160-3

Fischer-Baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: Evidence from intrusion errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(5), 1426–1446. doi: 10.1037/xlm0000102

Fischer-Baum, S., McCloskey, M., & Rapp, B. (2010). Representation of letter position in spelling: evidence from acquired dysgraphia. *Cognition*, *115*(3), 466–490. doi: 10.1016/j.cognition.2010.03.013

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), 458-67.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology. General*, *134*(4), 521-37. doi: 10.1037/0096-3445.134.4.521

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. doi: 10.1016/j.cognition.2010.07.005

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, *12*(4), 504–509. doi: 10.1111/j.1467-7687.2008.00794.x

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371. doi: 10.1016/j.cognition.2010.10.005

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362-6. doi: 10.1016/j.cognition.2010.01.006

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249-68.

Gervain, J., & Endress, A. D. (2017). Learning multiple rules simultaneously: affixes are more salient than reduplications. *Memory and Cognition*, *45*(3), 508–527.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: a study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287. doi: 10.1016/j.cognition.2012.06.010

Glicksohn, A., & Cohen, A. (2011). The role of gestalt grouping principles in visual statistical learning. *Attention, Perception and Psychophysics*, *73*(3), 708–713. doi: 10.3758/s13414-010-0084-4

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin and Review*, *11*(5), 791–806.

Graf-Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254-60. doi: 10.1111/j.1467-9280.2007.01885.x

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23-34.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. doi: 10.1016/j.cogpsych.2011.06.002

Henson, R. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive*

Psychology, *36*(2), 73-137.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, *12*(1), 131-41. doi: 10.1111/j.1467-7687.2008.00740.x

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339-45. doi: 10.1111/j.1467-7687.2009.00886.x

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*(9), 323-328.

Karaman, F., & Hay, J. F. (2018). The longevity of statistical learning: When infant memory decays, isolated words come to the rescue. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(2), 221–232. doi: 10.1037/xlm0000448

Kim, Y. J., & Sundara, M. (2015). Segmentation of vowel-initial words is facilitated by function words. *Journal of child language*, *42*, 709–733. doi: 10.1017/S0305000914000269

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*, 1399–1402.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(1), 79-91.

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and*

Language, *66*(1), 285 - 306. doi: 10.1016/j.jml.2011.09.004

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental science*, *14*, 1323–1329. doi: 10.1111/j.1467-7687.2011.01079.x

Makovski, T. (2016). Does proactive interference play a significant role in visual working memory tasks? *Journal of experimental psychology. Learning, memory, and cognition*, *42*, 1664–1672. doi: 10.1037/xlm0000262

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory and Cognition*, *36*(1), 43–52.

Marchetto, E., & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, *67*(3), 130 - 150. doi: 10.1016/j.cogpsych.2013.08.001

Marchetto, E., & Bonatti, L. L. (2015). Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities. *Journal of Child Language*, *42*(4), 873–902. doi: 10.1017/S0305000914000452

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199 - 227. doi: 10.1016/0024-3841(94)90342-5

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*(3), 303-315. doi: 10.1080/15475441.2011.609106

Miozzo, M., Petrova, A., Fischer-Baum, S., & Peressotti, F. (2016). Serial position encoding of signs. *Cognition*, *154*, 69–80. doi: 10.1016/j.cognition.2016.05.008

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology*, *28*(1), 28–38.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of personality and social psychology*, *35*(12), 847–862.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34. doi: 10.1111/j.1467-7687.2012.01189.x

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, *109*(3), 423 - 430. doi: 10.1016/j.cognition.2008.10.004

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750. doi: 10.1073/pnas.0708424105

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-7. doi: 10.1016/j.cognition.2009.07.011

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-7. doi: 10.1126/science.1072901

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299–1305. doi: 10.3758/MC.36.7.1299

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. doi: 10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of*

Memory and Language, *39*, 246–63.

Pilon, R. (1981). Segmentation of speech in a foreign language. *Journal of Psycholinguistic Research*, *10*(2), 113 - 122.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. doi: 10.1038/35107080

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-8.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74-85.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–21.

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language dependent cue for speech segmentation. *Developmental Science*, *20*(3). doi: 10.1111/desc.12390

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51-89.

Schonberg, C., Marcus, G. F., & Johnson, S. P. (2018). The roles of item repetition and position in infants' abstract rule learning. *Infant behavior & development*, *53*, 64–80. doi: 10.1016/j.infbeh.2018.08.003

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565-573. doi: 10.1111/j.1467-7687.2006.00534.x

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language*, *35*(1), 1-24.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: MIT Press.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193-247.

Shatzman, K. B., & McQueen, J. M. (2006a). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372-7. doi: 10.1111/j.1467-9280.2006.01714.x

Shatzman, K. B., & McQueen, J. M. (2006b). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, *68*(1), 1-16.

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in cognitive sciences*, *14*, 477–481. doi: 10.1016/j.tics.2010.07.002

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental science*, *11*, 407–413. doi: 10.1111/j.1467-7687.2008.00685.x

Shoaib, A., Wang, T., Hay, J. F., & Lany, J. (2018). Do infants learn words from statistics? evidence from english-learning infants hearing italian. *Cognitive Science*, *42*(8), 3083–

3099. doi: 10.1111/cogs.12673

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1-32. doi: 10.1016/j.cogpsych.2006.04.002

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6038–6043. doi: 10.1073/pnas.1017617108

Slone, L. K., & Johnson, S. (2015). Statistical and chunking processes in adults' visual sequence learning. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2218–2223). Austin, TX: Cognitive Science Society. Paper presented at the annual meeting of the cognitive science society.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92–102. doi: 10.1016/j.cognition.2018.05.016

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 1-11. doi: 10.1080/15475441.2015.1073153

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and cognitive processes*, *26*, 191–211. doi: 10.1080/01690961003787609

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cog-*

nitive Psychology, *50*(1), 86-132. doi: 10.1016/j.cogpsych.2004.06.001

Ter Schure, S., Mandell, D. J., Escudero, P., Raijmakers, M. E. J., & Johnson, S. P. (2014). Learning stimulus-location associations in 8- and 11-month-old infants: Multimodal versus unimodal information. *Infancy*, *19*, 476–495. doi: 10.1111/infa.12057

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, *67*(5), 867-75.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552-64. doi: 10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Van de Weijer, J. (1999). *Language input for word discovery* (MPI series in psycholinguistics, 9). Max Plank Institute for Psycholinguistics, Nijmegen.

van Alphen, P. M., & van Berkum, J. J. A. (2010). Is there pain in champagne? Semantic involvement of words within words during sense-making. *Journal of Cognitive Neuroscience*, *22*, 2618–2626. doi: 10.1162/jocn.2009.21336

van Heijningen, C. A. A., de Visser, J., Zuidema, W., & ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20538-43. doi: 10.1073/pnas.0908113106

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*, 375–382. doi: 10.1016/j.cognition.2013.02.015

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. doi: 10.1006/cogp.1995.1016

Weisstein, E. W. (n.d.). *Run. From MathWorld– A Wolfram Web Resource.* (Accessed on 07/13/2018)

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3-21.