# Words, rules, and mechanisms of language acquisition

Ansgar D. Endress
City University London, London, United Kingdom

Luca L. Bonatti
ICREA and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

Draft of September 21, 2015

We review recent artificial language learning studies, especially those following Endress and Bonatti (2007), suggesting that humans can deploy a variety of learning mechanisms to acquire artificial languages. Several experiments provide evidence for multiple learning mechanisms that can be deployed in fluent speech: one mechanism encodes the positions of syllables within words and can be used to extract generalization, while the other registers co-occurrence statistics of syllables and can be used to break a continuum into its components. We review dissociations between these mechanisms and their potential role in language acquisition. We then turn to recent criticisms of the multiple mechanisms hypothesis and show that they are inconsistent with the available data. Our results suggest that artificial and natural language learning is best understood by dissecting the underlying specialized learning abilities, and that these data provide a rare opportunity to link important language phenomena to basic psychological mechanisms.

## Introduction

Language acquisition requires a host of complex abilities that do not seem to be used in other domains (Lenneberg, 1967). Some of these abilities might require highly specialized and domain-specific mechanisms, while others might rely on more generic mechanisms such as statistical learning. Here, we review artificial grammar learning studies that have revealed dissociations between a variety of learning mechanisms, focusing on those experiments following Peña, Bonatti, Nespor, and Mehler (2002) and Endress and Bonatti (2007) because they present a relatively well worked out case of the kinds of information different learning mechanisms extract from speech, the underlying psychological mechanisms, as well as their role in the complex grammars of natural languages. We will then focus on some recent theoretical and computational proposals to account for these data without resorting to multiple mechanisms (Aslin & Newport, 2012; Laakso & Calvo, 2011; Perruchet, Tyler, Galland, & Peereman, 2004). We will explore how far they can succeed, and ultimately argue that they are not compatible with important aspects of the available evidence.

The research was supported by the Ministerio de Ciencia e Innovación Grant PSI2009-08232PSIC to L.L.B and PSI2012-32533 to A.D.E. as well as Marie Curie Incoming Fellowship 303163-COMINTENT. Further, we acknowledge support by grants CONSOLIDER-INGENIO-CDS-2007-00012 from the Spanish Ministerio de Economía y Competitividad and SGR-2009-1521 from the Catalan government.

### *The specificity of artificial language learning*

Language acquisition is a complex learning problem. We have only a limited understanding of the input infants face, which aspects of this input they can process, the mechanisms they use to process the input, and even the end state of a mature representation of one's native language. To abstract away from the complexities of natural languages, several researchers devised highly simplified artificial languages that mirror key characteristics of natural languages, but that allow investigators to monitor the processes of language acquisition in the laboratory.

By far the most influential artificial language studies concern the question of how words are extracted from fluent speech. To acquire words, infants need to know where they start and where they end, even though fluent speech does not contain the equivalent of white space in written language. It has long been assumed that there are no language-universal speech cues to word boundaries (but see Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Pilon, 1981). However, in a series of seminal demonstrations, Saffran and collaborators showed that even young infants are sensitive to distributional cues to word boundaries (Aslin, Saffran, & Newport, 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996), and suggested that infants might learn words from fluent speech by tracking transitional probabilities (TPs) among syllables. The underlying intuition is that syllables that are part of the

same word are more likely to occur together than syllables that are part of different words (Aslin et al., 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; see Batchelder, 2002; Brent & Cartwright, 1996; Perruchet & Vinter, 1998; Swingley, 2005, for related models implementing similar ideas). By now, there is overwhelming evidence that human infants and other animals can track TPs in speech and other stimuli (e.g., Aslin et al., 1998; Creel, Newport, & Aslin, 2004; Endress, 2010; Hauser, Newport, & Aslin, 2001; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran & Griepentrog, 2001; Toro & Trobalón, 2005) and, consequently, a near-universal consensus that infants can use TPs to extract words from fluent speech (although it is not clear how exactly TPs are used; see Endress & Langus, under revieww; Endress & Mehler, 2009b; Marchetto & Bonatti, 2015; Yang, 2004).

However, infants do not have to learn just the words of their native language. They also have to acquire its grammar. In the wake of these demonstration of the impressive statistical learning abilities in humans and other animals, different authors suggested that the very same abilities might also be used to learn much more abstract, grammatical features of language (e.g., Aslin & Newport, 2012; Bates & Elman, 1996; Elman et al., 1996; Saffran, 2001; Saffran & Wilson, 2003; Seidenberg, 1997), potentially showing that much of the innate and human-specific computational machinery previously supposed to be necessary for language acquisition (e.g., Lenneberg, 1967; Chomsky, 1975; Mehler & Dupoux, 1990) might be unnecessary. Considerable debates followed, trying to assess whether the statistical mechanisms that might be used to learn words from fluent speech might also be used for learning grammar (e.g., Bates & Elman, 1996; Elman et al., 1996; McClelland & Patterson, 2002; Saffran, 2001; Saffran & Wilson, 2003; Seidenberg, 1997; Seidenberg & Elman, 1999) or whether words and rules are learned using different mechanisms (e.g., Fodor & Pylyshyn, 1988; Marcus, Vijayan, Rao, & Vishton, 1999; Marcus, 1998; Pinker, 1999; Pinker & Prince, 1988; Pinker & Ullman, 2002; Peña et al., 2002; Endress & Bonatti, 2007), possibly relying on different cues contained in the input (Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Bonatti, Nespor, & Mehler, 2008; Toro, Shukla, Nespor, & Endress, 2008).

Against this background, Peña et al. (2002) provided a case where both word-learning and rule-learning could be observed simultaneously, and seemed to obey different constraints. (Here, we use rule-learning as a shorthand for the generalizations we will describe below, where will also discuss the specific underlying mechanisms.) These authors showed that, when human adults exposed to a sequence of trisyllabic "words," they can track TPs between non-adjacent syllables in a speech stream.[1] However, participants could not use these abilities for extracting certain rule-like "generalizations" within the words. These generalizations required additional cues, that could be as subtle as separating words by silence of just 25 ms. According to Peña et al. (2002), these silences may have provided segmentation cues, probably adding to the stream a minimal form of prosody, that allowed them to capture the generalizations hidden in the speech stream. Once even minimally exposed to such subliminally segmented streams, participants could extract generalizations that escaped them after listening to the same syllable sequences, but deprived of segmentation indexes. These data, as well as Endress and Bonatti's (2007) and Endress and Mehler's (2009a) results later on, provided an important case suggesting that certain simple grammar-like regularities could not be learned based on general statistical mechanisms alone, but rather required different specialized mechanisms, a conclusion that Endress and Bonatti (2007) dubbed the MOM (More than One Mechanisms) hypothesis: A statistical mechanisms might track TPs among adjacent and non-adjacent syllables, irrespective of whether silences are inserted between the words. When short silences are inserted between words, or other (prosodic) markers are given, a second mechanism allows participants to extract simple rule-like generalizations involving syllables that occur word-initially and word-finally, respectively, as if participants had learned a set of legal prefixes and a set of legal suffixes. This conclusion is reminiscent of Pinker's (1991, 1999) venerable proposal that mastery of words and rules in natural languages requires representations of a different nature; in our case, however, the use of artificial languages makes it easier to dissect the inner workings of the different mechanisms involved in extracting such representations, although it is still not clear to what extent artificial words and rules map onto actual words and rules in natural languages.

In the following, we will refer to these mechanisms as the statistical mechanisms and as the rule mechanism, respectively. Crucially, however, we do not rule out that each of these mechanisms might turn out to be composed of a collection of independent sub-mechanisms. In fact, we believe that it is most likely the case. Further, we do not claim that these two (collections of) mechanisms are the only mechanisms involved in language acquisition. Rather, we assume that language acquisition requires a multitude of mechanisms, of which these two might be a part; here, we focus on two specific mechanisms involved in the experiments reviewed below (see discussion). Here, we use these terms simply as short-hands for the mechanisms underlying performance with differ-

[1] Whilst Peña et al.'s (2002) conclusion that participants can track TPs between non-adjacent syllables has been initially challenged (Newport & Aslin, 2004; Onnis, Monaghan, Richmond, & Chater, 2005; Perruchet et al., 2004), it is now fairly accepted that participants are sensitive to such TPs also when various potential confounds are controlled for (Peña et al., 2002; Endress & Bonatti, 2007; Endress & Mehler, 2009a; Endress & Wood, 2011; Onnis et al., 2005).

ent types of test items, without prejudging their number or properties.

Unsurprisingly, the MOM hypothesis came under critical scrutiny. Two recent lines of criticisms are of particular importance. According to one of these lines of arguments, one single statistical mechanism might be sufficient to both explain word- and rule-learning once the "saliency" of certain representations is taken into account (Aslin & Newport, 2012). A second line argues on computational grounds that a Simple Recurrent Network (hereafter SRN) can reproduce the results presented in Endress and Bonatti (2007), allegedly showing that both rule- and word-learning can be accomplished by one single general-purpose mechanism (e.g., Laakso & Calvo, 2011).

We first review the evidence that led to the conclusion that different mechanisms are used to extract certain rule-like regularities in fluent speech and to track the statistical syllable distribution. Following this, we will show that Aslin and Newport's (2012) "saliency" hypothesis is insufficient to account for the data. We will then focus on one worked-out single-mechanism model (Laakso & Calvo, 2011), showing that it is contradicted by existing data and is incompatible with a series of well established psychological facts. We will conclude by exploring hypotheses on how the two mechanisms can be grounded in basic psychological mechanisms, and suggest that this line of investigation has the potential to link basic psychological processes to basic linguistic phenomena.

## A test case for word and rule-learning

### *An overview of the MOM hypothesis*

In Peña et al.'s (2002) and Endress and Bonatti's (2007) experiments, participants were familiarized with speech streams composed of trisyllabic words. The first and the last syllable of each word came from three possible "frames" while three different syllables could occur in the middle position. That is, words had the form $A_iX_kC_i$, where three $A_i\_C_i$ frames were combined with three $X_k$ middle syllables, yielding a total of nine words. In some experiments, the familiarization stream was continuous, with no silences between words; in other experiments, words were separated by 25 ms silences. Results showed that participants learned the TPs among syllables both with continuous and segmented familiarization streams. However, they learned that $A$ syllables had to occur word-initially and $C$ syllables word-finally only when segmentation cues were given.

These issues were tested using three critical types of test items among which participants had to choose after familiarization: "class-words," "part-words" and "rule-words." Class-words had the structure $A_iX'C_j$. That is, their initial and final syllables had occurred in these positions during familiarization, but never in the same word because they came from different frames. In contrast, their middle syllables had never occurred in the middle position during familiarization, but were $A$ or $C$ syllables. In other words, class-words had "correct" initial and final syllables, but had never occurred in the familiarization stream and had TPs of zero between all syllables.

Part-words had occurred in the speech streams, but straddled a word-boundary. That is, such words had either the structure $C_iA_jX$, taking the last syllable from one word and the first two syllables from the next word, or the structure $XC_iA_j$, taking the last two syllables from one word and the first syllable of the next words. Hence, part-words had occurred in the speech stream and had, therefore, positive TPs between their syllables, but they had "incorrect" initial and final syllables, and, therefore, incorrect "affixes."

Rule-words were like class-words, except that the first and the last syllable came from the same frame, yielding the structure $A_iX'C_i$; hence, they had "correct" initial and final syllables, and TPs of 1.0 between their first and their last syllable.

Given these items, it is possible to test whether participants track TPs among syllables, and whether, when segmentation cues are given, they simultaneously track the syllables at the word-edges (i.e., in the first and the last position). To test whether participants are sensitive to TPs between non-adjacent items, Peña et al. (2002) asked them to choose between words and part-words after exposure to a continuous stream. Because both items had similar TPs among adjacent syllables, but different TPs among non-adjacent syllables, a preference for words over part-words would suggest that participants tracked relations between nonadjacent syllables. After 10 min exposure, participants succeeded in the task. Further confirmation comes from experiments by Endress and Bonatti (2007) who asked participants to choose between rule-words and class-words. Given that rule-words and class-words are identical except that the first and the last syllable comes from the same frame in rule-words but not in class-words, participants should prefer rule-words to class-words if they had learned this statistical dependency between the first and the last syllable. Results showed that they did so both after continuous and after segmented familiarizations, suggesting that they could track TPs among syllables irrespectively of the presence of segmentation markers.

To test whether participants are sensitive to structural information in the speech stream, and whether they can learn "legal" prefix and suffix syllables, respectively, Endress and Bonatti (2007) asked them to choose between class-words and part-words. If they choose class-words, they must have learned the "legal" initial and final syllables; after all, the initial and final syllables are the only feature that class-words share with what participants had heard during the familiarization stream. In contrast, part-words have occurred during the familiarization and have, therefore, non-zero TPs.

Results showed that participants preferred class-words to part-words only when familiarized with a segmented stream, but not when familiarized with a continuous stream, suggesting that the segmentation cues were required to track the syllables at word edges. Moreover, Endress and Mehler (2009a) showed that participants specifically track information about the edges of words (i.e., their first and the last syllables) rather than arbitrary syllable positions. Using longer, five-syllable words (as opposed to the tri-syllabic words used by Endress & Bonatti, 2007), they showed that the generalizations are available only when the crucial syllables are the first and the last one (i.e., in words of the form $A_iXYZC_i$, where $A_i$ and $C_i$ are the critical syllables), but not when the crucial syllables are word-medial (i.e., in words of the form $XA_iYC_iZ$).

These results generalize the findings by Peña et al. (2002), who asked participants to choose between rule-words and part-words, either after exposure to a continuous stream or to a segmented stream. Compared to class-words, rule-words have TPs of 1.0 between their first and their last syllable, in addition to having legal initial and final syllables. Yet, participants preferred rule-words to part-words only when exposed to a brief segmented stream, and failed to do so after exposure to a continuous stream. Furthermore, receiving more familiarization did not help: participants always failed to find structural information when exposed to continuous streams, and even preferred part-words to rule-words after a 30 min familiarization.

Strikingly, also participants in Endress and Bonatti's (2007) experiments preferred class-words to part-words after short, segmented familiarizations, but this preference disappeared after 30 min of familiarization. Moreover, after a 60-min familiarization, participants even preferred part-words to class-words, reversing their initial preference. Hence, the rule-like regularities are available very quickly, whereas statistical information appears to require time and exposure to consolidate.

### *The evidence for multiple learning mechanisms*

Based on the results reviewed so far, Endress and Bonatti (2007) and Endress and Mehler (2009a) suggested that the rule-like generalizations and the statistical analysis of the speech stream relied on distinct mechanisms. Following up on these initial results, a variety of further investigations supports this view.

First, Peña et al. (2002), Endress and Bonatti (2007) and Endress and Mehler (2009a) showed that participants can track statistical regularities when exposed to a continuous speech streams, while the generalizations require segmentation cues. The fact that both kinds of computations require different kinds of cues would be unexpected if both relied on the same mechanism.

Second, the time course of both mechanisms seems to be fundamentally different. After short familiarizations, items conforming to the rule-like regularity are preferred to items that do not conform to these generalizations but that have stronger TPs; after very long familiarizations, this pattern reverses.

Third, the generalizations in artificial language experiments are observed predominantly when the crucial syllables are at the edges of words, while TPs are tracked fairly well when the critical syllables appear in word-medial positions. However, it seems problematic to postulate that a single mechanism both fails in non-edge positions (for generalizations), and succeeds in such positions (for TPs).

Fourth, the two mechanisms seem to have a different developmental time course. For example, when 18-months-old infants are exposed to artificial streams patterned after Peña et al.'s (2002) stimuli, but containing a conflict between statistical information and generalizations, they behave like adults. That is, when exposed to a continuous speech stream, they can extract statistically coherent items, but do not generalize structural regularities. In contrast, when exposed to a segmented speech stream, they generalize structural regularities, and choose them over statistically coherent items, again just like adults. Instead, 12-months-olds show a strikingly different pattern. Like adults and 18-month-olds, they can generalize structural generalizations when exposed to a segmented speech stream. However, they are unable to identify statistically coherent items when exposed to a continuous stream, even if this stream contains only minimal conflicts between statistical and structural information (Marchetto & Bonatti, 2013).[2] Hence, the ability to draw structural generalizations and to extract statistical information (across non-adjacent syllables) seem to arise at different ages, which seems difficult to reconcile with the view that both abilities rely on the same mechanism.

Fifth, the details of what the computations encode when acquiring a rule or when computing TPs seem different. Specifically, TPs and the rule mechanism behave in qualitatively different ways under temporal reversal. This fact has been shown by Endress and Wood (2011), who replicated Endress and Bonatti's (2007) and Endress and Mehler's (2009a) results with movement sequences (rather than speech material). Reproducing earlier results by Turk-Browne and Scholl (2009), they showed that participants are as good at discriminating high-TP items from low-TP items when these are played forward as when they are played backward. That is, if *ABC* is a high-TP item and *DEF* is a low-TP item, participants are as good if tested on *ABC vs. DEF* as when

[2] These results do not contradict the view that infants are sensitive to statistical information. While prior demonstrations of statistical learning in infants used statistical relations between adjacent syllables, Marchetto and Bonatti's (2013) experiments relied on statistical relations among non-adjacent syllables, and such relations are likely to be more difficult to track.

tested on *CBA vs. FED*.[3] In contrast, participants do not retain positional information when the test items are reversed, and never chose generalization items that are played backwards. Hence, TPs and the rule mechanism behave in qualitatively different ways under temporal reversal.

Sixth, the two mechanisms seem to encode spatial properties differently. Endress and Wood (2011) familiarized participants with a sequence of movements performed by an actor in frontal view. During test, however, the actor was rotated by $90^o$. While participants retained some sensitivity to rule-like generalizations after the actor had been rotated, they failed to discriminate high-TP from low-TP items. In other words, TPs and positional information appear to behave differently under spatial rotation. This result can be explained if these two mechanisms are independent, but it is harder to explain if they rely on the same TP-based mechanism.

Seventh, the brain mechanisms underlying rule- and word-extraction seem to be distinct. For example, using material closely patterned upon Peña et al.'s (2002), different authors suggested that ERP differ according to whether participants extract words or rules from the same speech stream (de Diego Balaguer, Toro, Rodriguez-Fornells, & Bachoud-Lévi, 2007; Mueller, Bahlmann, & Friederici, 2008). The learning of statistical regularities appeared correlated with a central N400 component, whereas the extraction of structural information was associated with an earlier P2 component (see also Mueller et al., 2008). Further, using speech streams similar to those used by Peña et al. (2002) , de Diego-Balaguer, Fuentemilla, and Rodriguez-Fornells (2011) suggested that statistical learning and the extraction of structural information are characterized by different patterns of dynamical brain activity. Long-range coherence between different regions of the scalp was found in different frequency bands for learning the statistical regularities and the structural regularities, respectively.

Eighth, results from the word segmentation literature also suggest that TP computations operate independently of the mechanisms that presumably underlie the generalizations, and are differentially accessible depending on the modality in which they are tested. Shukla, Nespor, and Mehler (2007) auditorily presented participants with "words" with high TPs that were either at the edges of an intonational contour or straddled a contour boundary. When tested with auditory test items, participants sometimes even preferred *foils* to the high-TP items that straddled the contour boundaries (see their Experiment 6, and Shukla, 2006, for more data), rejecting items straddling segmentation cues. At first sight, these results seem to show that prosody overrides TP computations. However, when participants were exposed to the same auditory familiarization but tested in the visual modality with written test items (rather than with auditory items), participants preferred high-TP items to the foils, reversing the preference Shukla et al. (2007) observed with an auditory test phase.[4]

Shukla et al.'s (2007) results suggest two crucial conclusions. First, participants must have computed TPs among syllables irrespective of whether the syllables straddled a contour boundary; had they not, they could not have preferred high-TP items in a visual test. Hence, the TP computations appear independent of, and unaffected by, the computations induced by boundary cues, directly contradicting single-mechanisms models. Second, a different mechanism must have led participants to *reject* the high-TP items when tested on auditory material and, as we will argue below, this mechanism is likely the same mechanism that led participants to generalize the structural regularities in the experiments reviewed above. (While Shukla et al. (2007) explained their results in terms of the alignment of words with prosodic boundaries, we will show below that our interpretation does not differ from theirs.)

In sum, beyond Peña et al.'s (2002) and Endress and Bonatti's (2007) initial results, there is considerable evidence for dissociations between rule mechanisms and statistical mechanisms. These dissociations relate to the cues used by either mechanism, their respective time courses of operation, the conditions under which they break down, their respective sensitivity to temporal order, their respective resilience to spatial rotation, their ontogenetic development, their brain mechanisms, the specificity of representations they create, and modality differences regarding how items can recognized.

In spite of this evidence, the conclusion that independent mechanisms might be at the root of these dissociations has recently been questioned (e.g., Aslin & Newport, 2012; Laakso & Calvo, 2011; Perruchet et al., 2004). We will now critically examine these proposals, using the evidence reviewed above as a yardstick to assess the viability of such alternatives.

## Current alternatives to the MOM hypothesis

### *Chunking and generalizations*

Perruchet et al. (2004) proposed that their PARSER model (Perruchet & Vinter, 1998) provided a better account of some data, and notably of Peña et al.'s (2002)

[3] While these results seem to suggest that TPs are not directional, Turk-Browne and Scholl (2009) also showed that forward items can be discriminated from backward items, that is, participants prefer *ABC* to *CBA*, suggesting that TPs retain some directional information as well.

[4] Shukla (2006) also showed that prosodic segmentation cues are not just one of the cues over which TP-like co-occurrence statistics can be computed. They controlled for this possibility by asking whether participants would recognize test items that contained exactly the same prosodic properties as during familiarization; results showed that they did not. Shukla et al. (2007) also showed that the modality of presentation counts.

results. However, it is easy to see that their model would be unable to account for any kind of generalization. PARSER operates by recursively chunking units from a speech stream. For example, if it encounters A followed by B, it might create a unit AB. If, at a later point, the unit AB is followed by C, a new unit ABC might be created. Recurring units are strengthened, while spurious units are eliminated through decay and interference. The fact that PARSER cannot account for generalizations can be proved. Assume that PARSER creates a unit for a generalization item $XYZ$ that conforms to a regularity, but has not been encountered in the speech stream (e.g., a class-word of the form $A_iX'C_j$). If PARSER accepts the item $XYZ$, then $XY$ must have been followed by $Z$, or $X$ must have been followed by $YZ$. In either case, the sequence $XYZ$ would be attested in the speech stream, which contradicts that $XYZ$ is a generalization item. Hence, PARSER cannot accept class-word like generalizations. This proof by contradiction shows that PARSER might or might not account for word-segmentation results, but not for any kind of generalization. As a result, we will not discuss it further.

### *Saliency, statistics and the saliency of the available evidence*

Aslin and Newport (2012) argue that a single learning mechanism enhanced by some measure of the salience of the stimuli can circumvent the difficulties faced by a pure single mechanisms hypothesis. Indeed, Aslin and Newport (2012) acknowledge that one single mechanism, unaided by other factors, cannot account for many aspects of language acquisition. However, they argue, such a mechanism would normally not operate on raw sensory input. Instead, different parts of the input carry different saliency, which, in turn, might provide a more structured input on which statistical processes can operate. According to these authors, the saliency of a representation originates from diverse sources. In some cases, it comes from Nature. In other cases, it may follow from a small set of patterns shared by all languages of the world. In still other cases, it can be modulated by context: when Nature does not help, context can signal that a pattern is relevant for either rule acquisition or word acquisition.

Thus, the core idea of this theory is that, when stimulus dimensions or other factors make one aspect of a stimulus salient, learners will generalize to novel stimuli sharing these salient features by virtue of the same statistical mechanism that, in the absence of saliency, would account for word segmentation. For example, when the computations apply to a monotonous sequence of syllables, the mechanism will compute transitional probabilities among syllables. When, instead, the input to the statistical mechanism is composed of pre-analyzed representations that attract the learner's focus to some stimulus features (e.g., salient edge syllables of words), participants will accept novel stimuli that have the same salient features, and, therefore, behave as if they had formed generalizations. As a result, rule acquisition and word segmentation "are in fact not distinct, but rather are different outcomes of the same learning mechanism" (p. 172).

We agree with Aslin and Newport (2012) that saliency might play a crucial role in language processing and acquisition, and some of us have argued that the saliency is crucial to determine which rule can be easily acquired (e.g., Endress, Scholl, & Mehler, 2005; Endress, Dehaene-Lambertz, & Mehler, 2007). However, we do not believe that Aslin and Newport's (2012) "statistics+saliency" theory is sufficient to account for the data presented above, let alone for language acquisition.

Under one interpretation, the claim that some regularities are more salient than others is simply a different way to say that different regularities are learned by different mechanisms, and that these mechanisms make some regularities more salient than others. As such, this hypothesis is consistent with the data, because all applicable non-statistical mechanisms are simply summarized under the label "saliency." In fact, it is well known that statistical computations over suitably rich representations might have a powerful role in language acquisition (e.g., Yang, 2010); however, simply labeling such rich linguistic representations as "salient" does not explain what makes them so rich and so salient.

Under another interpretation, however, the hypothesis falls short of explaining many facts about language acquisition, even if we restrict our analysis to the simple data on artificial language learning such as the data reviewed above.

For example, there is substantial evidence that participants extract words, but not rules, by preferentially relying on consonants, and extract rules, but not words, by preferentially relying on vowels (Bonatti et al., 2005; Toro, Bonatti, et al., 2008; Toro, Shukla, et al., 2008). In these experiments, the statistical relations among vowels are identical to those among consonants. As a result, a single mechanism account of these results would be forced to postulating that this dissociation occurs because consonants are more salient for word-extraction while vowels are more salient for rule-extraction, even though word-extraction and rule-extraction supposedly rely on the same mechanism. However, this solution clearly begs the question of why the same mechanism sometimes finds vowels more salient, and sometimes consonants. One possibility is that the different behaviors of vowels and consonants might be due to prior statistical computations performed over the entire linguistic input a speaker has received (see e.g. Keidel, Jenison, Kluender, & Seidenberg, 2007, but see Bonatti, Peña, Nespor, & Mehler, 2007; Toro, Shukla, et al., 2008, for discussion). However, when one considers the available evidence such an account becomes implausible. For example, it fails to explain why vowels are particularly

salient for rules and consonant for words, and not vice-versa (Bonatti et al., 2007). Likewise, it fails to explain why differential roles of vowels or consonants have also been found in lexical access during reading (New, Araújo, & Nazzi, 2008), in word learning tasks with two and three year-old children (Havy, Bertoncini, & Nazzi, 2011; Nazzi, 2005) and even in prelinguistic infants (Hochmann, Benavides-Varela, Nespor, & Mehler, 2011; Pons & Toro, 2010), and why the cerebral representation associated with consonants and vowels are dissociable (e.g., Caramazza, Chialant, Capasso, & Miceli, 2000; Carreiras & Price, 2008).

A second example is provided by Peña et al.'s (2002) results. As mentioned above, these authors showed that participants switch from TP computations to rule-learning on the basis of subtle subliminal segmentation cues present in a stream. A "statistics+saliency" hypothesis could explain this result roughly in the following way. Without segmentation cues, the (statistical) mechanism simply computes TPs among syllables. In contrast, when such cues are present, edge positions become salient, and the statistical computations will be influenced by the edges' saliency, and preferentially operate over those salient positions. However, it is not at all clear that a statistical model enhanced by saliency information would end up learning structural information from a stream. To see why this is the case, consider what saliency means. Presumably, the representation of a salient item is more conspicuous, such that it becomes more active. However, according to many models of associative learning, associations between more active items are stronger, because the weight changes are proportional to the activation of the items entering the associations.

In Endress and Bonatti's (2007) experiments, the items that become more salient due to the segmentation cues are clearly the first and the last syllable in a word (i.e., the $A$ and the $C$ syllable). Hence, both within-word and between-word associations among these syllables should be strengthened when segmentation cues are given. In Endress and Bonatti's (2007) crucial test items, both $A$ and $C$ syllables occur, albeit from different frames. Part-words contain $CA$ between-word transitions; further, they contain transitions between the salient syllables and the middle syllables from the stream (i.e., part-words have the form $XCA$ or $CAX$). In contrast, in class-words, having the form $A_iX'C_j$, the order of the salient syllables is reversed relative to familiarization; that is, in a between-word transition between $C_j$ and $A_i$, $C_j$ preceded $A_i$, while the opposite is the true for class-words. Moreover, the salient syllables are more distant in class-words than in the between-word transitions.

Given this pattern of associations, one would expect the preference for part-words over class-words to be strengthened when segmentation cues are given; hence, if the role of the segmentation cues were really to make certain syllables more salient, participants should increase their preference for part-words to class-words when segmentation cues are given as opposed to when familiarized with a continuous stream. In contrast to this prediction, participants preferred class-words to part-words with segmented familiarizations, but not with continuous familiarizations, showing exactly the opposite behavior. The role of the edges thus does not seem to be to make syllables salient, but rather that words have to be *aligned* with the edges which, in turn, enables generalizations (see also Shukla et al., 2007).

Of course, much depends on the exact implementation of the "statistics+saliency" account, and on what exactly is meant by saliency. Our point, however, is that it is far from clear that adding saliency to statistical computations helps a single-mechanism theory to explain the data.

### *Simulating rule acquisition with a single statistical mechanism*

A key question for theories about mechanisms involved in language acquisition is to assess to what extent a single-mechanism, associative model can explain the complex behavior of learners in both artificial and natural language acquisition settings. Endress and Bonatti (2007) explored this question by examining a prominent candidate single-mechanism model that has been widely used in the study of language acquisition: a Simple Recurrent Network (SRN; Elman, 1990). Its basic idea is that the network is trained to predict the next syllable in the speech stream based on the previous syllable(s).

Endress and Bonatti (2007) tested a large number of network parameters, as well as training conditions. Their results clearly showed that the overall pattern of the simulations was not compatible with the behavioral data. However, they also found that, under very specific assumptions, the networks seemed to reproduce the preference for class-words over part-words. The trouble is, as Endress and Bonatti (2007) showed, that such assumption are all problematic and sharply differ from humans' actual behavior , suggesting that, overall, a single-mechanism account of the data was not psychologically plausible.

In contrast, Laakso and Calvo (2011) presented a detailed case in favor of a single-mechanism explanation of Endress and Bonatti's (2007) results. Laakso and Calvo (2011) replicated a subset of Endress and Bonatti's (2007) simulations, where they made the problematic assumptions identified by Endress and Bonatti (2007) but studied a slightly different network, with more hidden units and a different activation function.

As expected, Laakso and Calvo (2011) mostly replicated Endress and Bonatti's (2007) simulations. However, the authors drew markedly different conclusions, interpreting their results as supporting a single-mechanism theory of language acquisition. However, as we will show below, this conclusion relies on selectively ignoring simulation results, and ignoring many empirical

phenomena that would contract this account.

*Discrepancies between the experiments modeled and the empirical data.* Laakso and Calvo's (2011) crucial argument relies on the claim that, with segmented familiarizations, the network prefers class-words to part-words after few training cycles, and part-words to class-words after more familiarization cycles. However, as is clear from their Figure 4, and as Endress and Bonatti (2007) already observed (p. 283), the network reverses the preference only against part-words with the structure $C_iA_jX$, while the network prefers class-words to part-words with the structure $XC_iA_j$ after all numbers of training cycles. In contrast, with human data, Endress and Bonatti (2007) did not find such an asymmetry between part-word types.

Endress and Bonatti (2007) already commented that the reason for the asymmetry in how the network performed on the two part-word types "lies in a quirk of the representation induced by the familiarization onto the network that does not seem to affect participants. When silences are represented as extra-symbols during familiarization, the network learns that a silence follows a 'C' syllable with certainty. During the test phase, because the second syllables of part-words of type $[XC_iA_j]$ are precisely 'C' syllables, the network will systematically predict an incorrect syllable, unless the silences are also included in the part-words" (p. 283).[5]

Not only does the model makes incorrect prediction about the dynamics of when some items should be preferred to others, but it also wrongly predicts the strength of such preferences. For example, in Study 1, where the network was familiarized with a segmented stream, the model predicts that, after short familiarization durations, the preference for class-words over part-words should be much stronger than the preference for words over rule-words. As shown in Figure 1, using the values from Laakso and Calvo's (2011) Table B1, one obtains an effect size (Cohen's $d$) of 8.93 for the class-word vs. part-word discrimination, and of 1.18 for the word vs. rule-word discrimination. Human data show exactly the opposite pattern: the class-word vs. part-word discrimination yielded effect sizes of .64 in Endress and Bonatti's (2007) Experiments 3, while the word vs. rule-word discrimination yielded an effect size of 1.59 in Endress and Bonatti's (2007) Experiment 8. [6]

The network also makes a prediction that seems to contradict well-established principles of psychology. In Laakso and Calvo's (2011) Study 1, the preference for words over part-words follows an inverted U-shaped pattern when considering means, and shows decreasing performance when considering effect sizes corresponding to the discrimination (see Figure 2). This, however, contradicts basic findings in the psychology of memory. To see why this is the case, consider Endress and Mehler's (2009a) experiments, where words were presented in isolation, separated by silences of 1 s. Hence, the stream consisted of a clearly distinguishable sequence of words. Further, the subsequent two-alternative forced-choice task just amounts to a memory test for words.

Given that Laakso and Calvo (2011) propose, and we agree, that the 1-s separation is computationally equivalent to the 25-ms silences used by Endress and Bonatti (2007), their model predicts that memory for words should be worse when words are presented more often. However, Ebbinghaus (1885/1913), and many authors after him, have shown that presenting items more often helps memory performance and does not hurt it. That said, participants might not only learn the words themselves but also subsequences of the words (e.g., syllable pairs). As such, they might also become more familiar with part-words as items are presented more often (unless they learn to *reject* part-words because they straddle word-boundaries, in line with Shukla et al.'s (2007) results). However, it is quite implausible that it would become increasingly hard to discriminate between actual words and words that combine syllables from different memory items that do not respect their onsets and offsets. Given that Laakso and Calvo's (2011) model contradicts one of the best-established facts of experimental psychology, it seems plausible to conclude that their representational scheme has little relation with the actual human processing system.

*Can the model be extended to account for other evidence?.* As reviewed above, several lines of evidence following Endress and Bonatti's (2007) work have provided further support for the MOM hypothesis. Unfortunately, the existing single-mechanism model does not consider them. More importantly, it is unclear how any of its extensions could account for them. The model (1) does not explain why a single mechanism sometimes breaks down in word-internal positions and sometimes does not (Endress & Mehler, 2009a); a multiple

[5] Laakso and Calvo (2011) tried to explain away the apparent discrepancy between the data and their central result arguing that, after all, Endress and Bonatti's (2007) data were not so compelling because they lacked adequate statistical power, arguing that "a sufficiently powerful test of the hypothesis that participants [would] respond differently to part words of different types [was] therefore needed (p. 18)." On a general level, this criticism is certainly possible, just as it is possible that any statistically significant result is obtained by chance (albeit with low probability). However, Laakso and Calvo (2011) did not provide any evidence to support this ad-hoc argument, not even a power analysis to evaluate the hypothesis that the relevant experiments lack statistical power, nor do they show that the test they proposed addresses the alleged problem of statistical power in any way. Their claim thus appears unsupported.

[6] In Experiment 10, Endress and Bonatti (2007) used different stimuli than in Experiments 3 and 8; the resulting class-word vs. part-word discrimination yielded an effect size of 1.24, and thus does not show the marked advantage for the class-word vs. part-word discrimination shown by Laakso and Calvo's (2011) network.

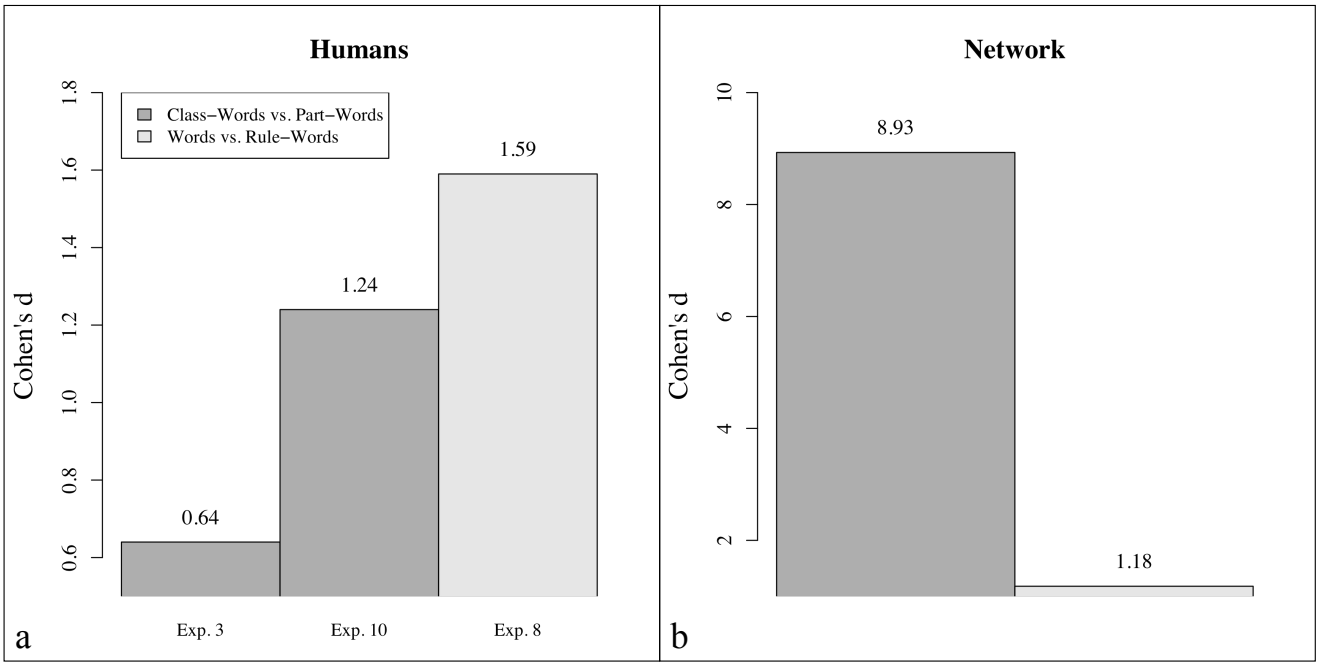


*Figure 1.* Effect sizes (Cohen's *d*) for the class-word vs. part-word discrimination (dark bars) and the word vs. rule-word discrimination with segmented 2-min streams in (a) humans and (b) networks. (a) In humans, the word vs. rule-word discrimination is numerically easier than the class-words vs. part-word discrimination. Note that Endress & Bonatti (2007) used a different stimulus materials in Experiment 10 than in Experiments 3 and 8. (b) After the number of training cycles Laakso & Calvo (2011) propose to correspond to a 2-min familiarization, the network performance on the class-word vs. part-word discrimination is much better than on the word vs. rule-word discrimination, showing the opposite pattern from humans.

mechanism model provides a natural account for such a dissociation, because one of the mechanisms might be operating in word-internal positions while the other one might not. Further, it does not explain (2) why one of the mechanisms appears to be sensitive to temporal order while the other one is not; (3) why one mechanism seems to provide viewpoint-invariant representations of sequences of actions, while the other one does not (Endress & Wood, 2011); (4) why one mechanisms appears to have different neural correlates from the other (de Diego Balaguer et al., 2007; de Diego-Balaguer et al., 2011; Mueller et al., 2008); and (5) why the mechanisms differ in how the information they extract can be recognized across modalities (Shukla et al., 2007). In the absence of a model that accounts for these dissociations, we believe that the most straightforward conclusions is that the underlying mechanisms are indeed distinct.

In addition to these newer lines of evidence, other data already presented by Peña et al. (2002), Endress and Bonatti (2007) and Endress and Mehler (2009a) seem to contradict the single-mechanism model.

For example, Peña et al. (2002) (Footnote 27) reported the following experiment. They familiarized participants with a segmented 10 min stream. Following this, they asked participants to choose between rule-words and part-words including 25 ms silences between the $C$ and the $A$ syllable; that is, these part-words had the structure $XC_i\#A_j$, where # stands for a 25 ms silence. Results showed that, just as in Peña et al.'s (2002) Experiment 3 where part-words did not contain silences, participants preferred rule-words to part-words. These results directly contradict the single-mechanism model. Given that that it is trained to predict syllables from silences, it would necessarily predict a stronger preference for part-words when these contain silences, simply because silence-containing part-words reflect *exactly* the statistical structure of the part-words in the speech stream.

Endress and Mehler (2009a) also provided problematic evidence for the model. As mentioned above, Endress and Mehler (2009a) used penta-syllabic words (as opposed to the trisyllabic items used by Endress & Bonatti, 2007). They showed that the positional generalizations can be performed when the critical syllables are in the first and the last position of words, but not when they are in the second and the fourth position: When familiarized with a segmented stream, participants preferred class-words to part-words, but only when the critical syllables were in the first and the last

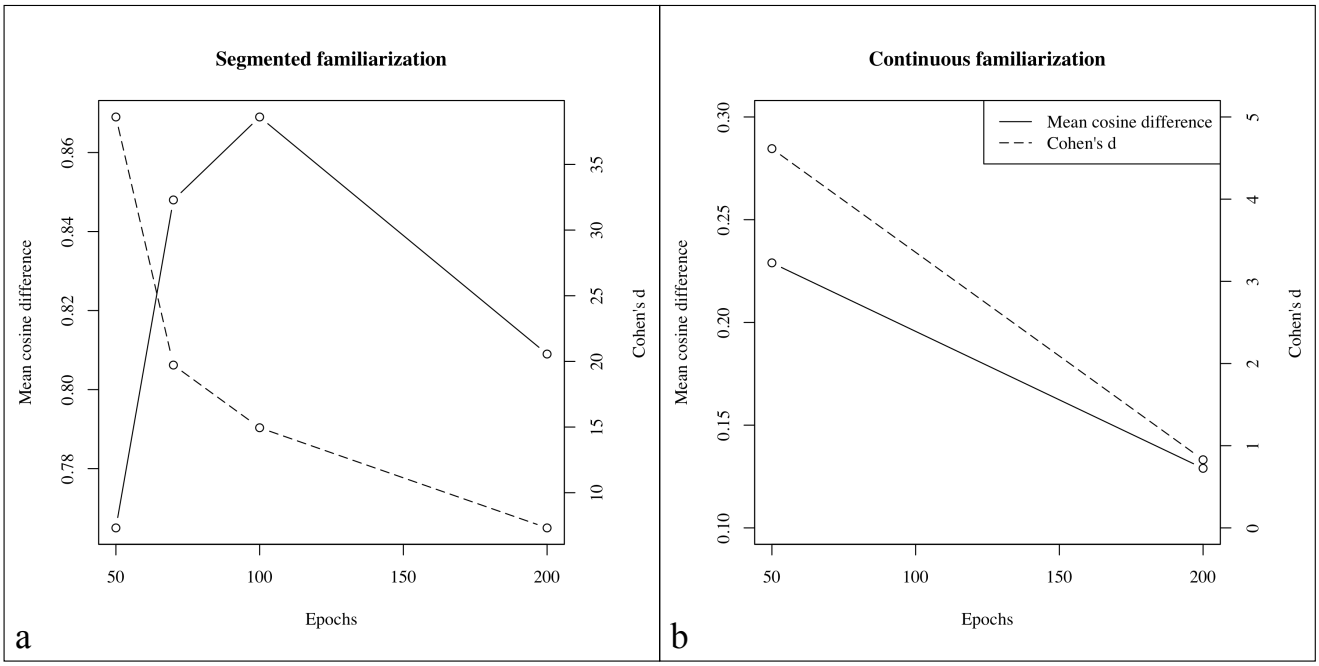


*Figure 2.* Difference in predictions for the last syllable of words and part-words, respectively, after (a) a familiarization with a segmented stream and (b) a familiarization with a continuous stream. The solid line shows the average difference between the cosinus values between the predicted network output and the target "syllables." The dashed line shows the corresponding effect sizes (Cohen's d). Laakso & Calvo (2011) simulations show that it should be harder to recognize words when they are encountered more often.

position, and not when the critical syllables were word-internal. In contrast, when familiarized with a continuous stream, participants preferred part-words to class-words, with no difference due to the location of the critical syllables.

It is difficult to see how the model can be extended to account for such results. Laakso and Calvo (2011) claim that they "can easily be accommodated within the general framework herewith advocated" (p. 30). However, if, as Laakso and Calvo (2011) assume, the generalizations are computed by associations between a single boundary marker (e.g., a symbol for the silences) and items in the critical positions within words, generalizations in the fourth position should be easier to track than in the last position, simply because the fourth position is closer the marker of the onset, and because it is well known that associations between closer items are easier to track than associations between more distant items (Ebbinghaus, 1885/1913). The single-mechanism model seems to suggest the contrary.

We verified this intuition by running simulations with an SRN, using 450 parameter sets. For each parameter set, we simulated an experiment with 20 participants.[7]

To reproduce the edge advantage for the generalizations, the network needs to exhibit (i) a significant preference for class-words to part-words in the edge condition; (ii) a (significant or non-significant) preference for part-words to class-words in the middle condition; and (iii) a significant difference (i.e., interaction) between the preference for class-words over part-words and the edge vs. middle manipulation. At least in the parameter set explored, there was not a single simulated experiment fulfilling the conditions. In fact, there was not a single simulation where class-words were preferred to part-words in the edge condition.

Further, in most simulations, the preference for class-words was at least numerically stronger in the middle condition than in the edge condition (see Figure 3).[8] Hence, an SRN does not easily account for Endress and

[7] We used the same network architecture as Endress and Bonatti (2007), except that, following Laakso and Calvo (2011), we used 54 hidden units and set the momentum to 0. We varied learning rates between $10^{-5}$ and .9 in 15 steps, and learning cycles between 10 and 300.

[8] Like Endress and Bonatti (2007) and Laakso and Calvo (2011), we exposed the network to a segmented stream, and then tested the network's preferences by recording its output for the target syllable of the test items, using the cosine similarity measure. However, in the middle condition, there are two ways to define the target syllable. Given that the critical syllables for the generalizations are in the second and the fourth position in the middle condition, the most appropriate choice for the target syllable is arguably the fourth syllable. Alternatively, one might also choose the last syllable. For

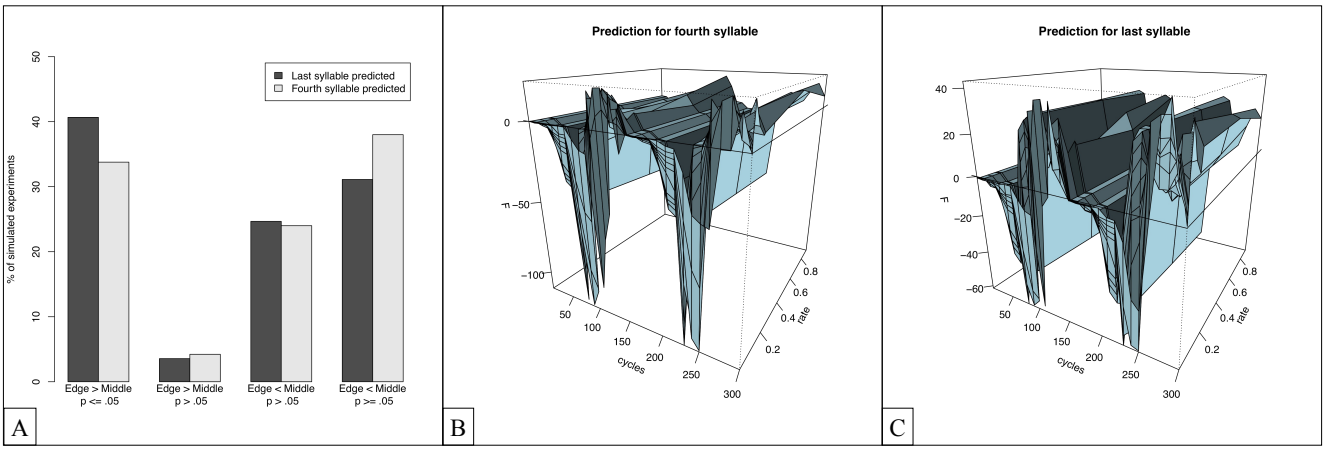


*Figure 3.* We simulated experiments by recording the results of 20 simulations with different network initializations, representing 20 participants. One experiment was simulated for each set of network parameters. The networks did not reproduce the preference for class-words over part-words in the edge condition for any set of network parameters. However, for completeness, we report more detailed results. (A) Proportion of simulated experiments where the preference for class-words over part-words is (significantly or numerically) stronger in the edge condition or the middle condition, depending on whether, in the middle condition, the fourth or the fifth syllable is considered as the target syllable. For most simulated experiments, the preference for class-words is stronger in the middle condition than in the edge condition (i.e., the preference for part-words is weaker), suggesting that the network does not intrinsically account for Endress & Mehler's (2009a) data. (B) F-values associated with the interaction between the preference for class-words over part-words and the edge vs. middle condition when the fourth syllable is considered the target syllable. When the preference for class-words was stronger in the middle condition, the F-values were multiplied by -1. (C) F-values associated with the aforementioned interaction when the fifth syllable is considered the target syllable.

Mehler's (2009a) results: the network fails to reproduce the basic psychological phenomenon that events in edges of sequences are easier to process than edge in sequence-middles.[9]

In sum, over and above the novel experiments that further supported the MOM hypothesis, Endress and Bonatti's (2007) work already provided evidence which seems to be incompatible with Laakso and Calvo's (2011) single-mechanism network. Of course, the model accounts for some aspects of the data. However, partial simulations of partial data are neither interesting nor useful *per se*, unless they are used to build more comprehensive models that do actually account for the data. It is certainly possible to construct a model that accounts for some aspects of the data, and another model that accounts for another aspect. But the question is whether there is a more general model that accounts for the general pattern of empirical data, and we see no easy way (assuming any exists) in which this can be done for the single-mechanism model we analyzed. Because this is the only existing model sufficiently detailed to allow critical investigation, we believe that theoretical conclusions against the existence of multiple mechanisms in language acquisition are at best premature.

## Mechanisms for the MOM hypothesis

We have argued that a considerable amount of psychological evidence suggests that several mechanisms are at the root of the performance in Endress and Bonatti's (2007) experiments, and, we surmise, in language acquisition in general. However, what might the underlying psychological mechanisms be? We will now outline a possible psychological model of these data, based on two well known memory mechanisms. Although tentative, our account provides a better theory than extant single mechanism accounts, in that it provides a natural explanation of most of the experiments reviewed above (and is at least compatible with the rest), is more parsimonious and grounded in basic aspects of memory processing.

completeness, we represent both possibilities in Figure 3

[9] While participants in Endress and Mehler's (2009a) experiments were not directly tested on their retention of class-words but rather had to choose between class-words and part-words, the statistical structure of the part-words as well as Endress and Mehler's (2009a) Experiment 2 suggest that there is no intrinsic preference for part-words depending on whether the crucial syllables are at the edges of words or word-internal. As a result, the preferential learning of the positional generalization when the critical syllables are in word-edges reflects better learning of sequential positions at word-edges.

Importantly, we will also suggest that this model may provide one of the rare cases where basic psychological processes can be linked to important phenomena in real languages.

## *Mechanisms for generalizations*

There are two basic facts that any model of the aforementioned data needs to explain. First, Peña et al. (2002) and Endress and Mehler (2009a) showed that participants can identify words in a speech stream by tracking TPs among adjacent and non-adjacent syllables. Second, once segmentation markers such as short silences are inserted between words, participants also become sensitive to rule-like generalizations within words, and track the syllables that occur word-initial and word-finally.

As it happens, the serial memory literature has revealed different kinds of memory encodings for sequences that seem to fit these facts, known as chaining memory and ordinal memory (we will refer to the latter type of memory as "positional" memory for consistency with Endress and Mehler's (2009a) terminology). Specifically, a sequence like *ABCD* might be encoded in two different ways (see e.g. Henson, 1998, for a review). First, people might encode it in terms of the actual transition between elements (e.g., $A \rightarrow B \rightarrow C \rightarrow D$), a coding scheme that is, at its root, a deterministic version of TPs. Second, people might encode it structurally, by reference to the positions of the sequence items, relative to the first and the last position (e.g., Conrad, 1960; Henson, 1998, 1999; Hicks, Hakes, & Young, 1966; Ng & Maybery, 2002; Schulz, 1955). They might know that $A$ came first, $D$ came last, and $B$ and $C$ occurred at some distance from the first and the last position. Endress and Bonatti (2007) and Endress and Mehler (2009a) suggested that two mechanisms involved in word learning and rule-like generalizations might be probabilistic versions of chaining memory and positional memory. This account differs only in one aspect from the aforementioned memory models: while it is generally assumed in the serial memory literature that participants have access to either one mechanisms or the other (see e.g. Henson, 1998, for a review), Endress and Bonatti (2007) and Endress and Mehler (2009a) suggested that participants might use both mechanisms simultaneously.

We suggest that participants' sensitivity to TPs might be due a mechanism akin to chaining memory. In contrast, the rule-like generalizations might rely on a mechanism akin to positional memory, and this mechanism might require segmentation cues to track structural aspects of the items, such as the first and the last syllables in words. We now show how this simple explanation offers a natural account of the most crucial available data.

## *Explaining the available data*

First, the hypothesis clarifies the role of the segmentation cues in the experiments we reviewed. Segmentation cues provide cues to edges of constituents, so that positional representations can be constructed. These representations, in turn, might allow participants to compute the generalizations. Importantly, and as discussed in more detail above, claiming that segmentation cues simply increase the saliency of some syllables is not sufficient to account for the generalizations. If it were, then combinations of edge syllables that straddle word boundaries (as in part-words) should be as good as syllables that occur in their correct — edge — position. Rather, as has long been proposed in linguistics (e.g., McCarthy & Prince, 1993; Nespor & Vogel, 1986), constituents have to be *aligned* with edges, and the positional memory mechanism provides a reason why this is so.

Second, it explains the negative correlation of the preference for the generalization items and the familiarization duration. According to this account, participants should be familiar with the positional generalizations early on, simply because edges, providing reference points for encoding positions, are salient elements of words. As a result, edge-based generalizations should be computed relatively quickly, which is just what the data suggest. In contrast, the alternative, TP-based choice can become familiar only by tracking the syllable distribution in the speech streams. Because such a distributional analysis presumably takes time, a natural prediction is that the familiarity with TPs should strengthen over time. As a result of these two computations, the preference for the generalizations should decrease with exposure. Our data show that it does.

Third, it explains why only statistical information, but not generalizations, can be tracked in backward items. While participants can track backward TPs (Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008; Turk-Browne & Scholl, 2009), potentially because TPs are not directional, generalization items should not be recognized after temporal reversal. Indeed, because reversing such items switches the first and the last position, the reversed items do not have "correct" initial and final movements, and participants have no reason to choose them.

Fourth, it offers a account of why participants reject statistically well-formed items that straddle prosodic boundaries when tested in the auditory modality, but why they accept such items when tested in the written modality (Shukla et al., 2007). Prosodic contour boundaries give edge-cues of the kind required for a positional memory encoding. This, in turn, might allow participants to reject items that do not have their syllables in the correct positions, because, as Shukla et al. (2007) argues, these words are not aligned with the edges. In contrast, as proposed by Shukla et al. (2007), "episodic information" (that we identify with the prosodic edge

information) cannot be accessed from written material, resulting in the dissociation between the visual and the auditory test modality.

Fifth, the hypothesis is consistent with Endress and Wood's (2011) finding that, in the visual modality, the generalizations tolerate a spatial rotation of the stimuli, while the statistical computations do not. Such results are hard to reconcile with a single-mechanism model for both generalizations and statistics; in a dual-mechanism model, in contrast, one of these mechanisms might well be more tolerant to spatial rotations than the other.

Last, but not least, it is consistent with the finding that rules and words are represented differently in the brain and have a different developmental course. This is exactly what one would expect if the mechanisms were different and dissociable.

It should be noted that not all of the above results could have been predicted on a priori grounds from the hypothesis that the generalizations exploit a positional memory mechanism that is distinct from the statistical mechanisms tracking TPs. However, in contrast to the alternative theories reviewed above, this account is at least consistent with these data.

### *Positional memories, co-occurrence statistics, and real languages*

Endress and Bonatti (2007) and Endress and Mehler (2009a) proposed that two mechanisms can be observed when humans analyze fluent speech. When cues to word boundaries are given, participants can track the syllables that occur at the edges of words, and can use this information to compute grammar-like generalizations. In contrast, irrespective of the presence of segmentation cues, participants compute TPs among (adjacent or non-adjacent) syllables. We have shown that this simple explanation has led to a series of predictions that have been tested experimentally.

It turns out that these mechanisms do not only provide natural accounts of numerous artificial language learning studies, but might also yield a psychological explanation for several crucial aspects of languages. That said, we certainly do not propose that language can be explained just based on memory mechanisms, or that the two mechanisms explored here are sufficient to explain all artificial language learning studies. Rather, we suggest that the two memory mechanisms discussed here have been recruited by language for linguistic purposes, but that other aspects of language rely on different mechanisms (see Endress, Cahill, Block, Watumull, & Hauser, 2009; Endress, Nespor, & Mehler, 2009, for discussion). Still, these two mechanisms allow us to link specific linguistic regularities to their underlying psychological mechanisms, and give us insight into their evolution.

Specifically, it is possible that words and, in fact, other constituents in natural languages are encoded by means of the positional memory mechanisms described above: that is, it is possible that the positions of the units within words (e.g., syllables) is also encoded relative to their edges. This hypothesis is supported by artificial language learning studies (Endress & Mehler, 2009b; Endress & Langus, under revieww; but see Perruchet & Poulin-Charronnat, 2012) and by the types of errors brain-damaged patients make with written words (Fischer-Baum, McCloskey, & Rapp, 2010). The same hypothesis has been used to develop some formal treatments of various language phenomena (McCarthy & Prince, 1993).

To give a simple example, among the languages of the world that have fixed stress, all assign stress relative to word-edges (e.g., Goedemans & van der Hulst, 2008a; Hayes, 1995). That is, not all languages have word-initial or word-final stress (e.g., Italian has mostly penultimate stress), but stress is located either in the first three or in the last three positions, and the positions are counted from the edges. Even in languages where the stress location is not fixed but depends on other phonological and morphological factors, the stress location is often determined relative to the first and the last position (Goedemans & van der Hulst, 2008b). This cross-linguistic regularity directly follows from the hypothesis that words are encoded using a positional memory mechanism: given that speakers need to determine the position of the stressed syllable, they do so in positions that can be tracked easily, and these are the positions close to the first and the last syllable. In contrast, such results are hard to explain by a TP-like mechanism, simply because such mechanisms do not have a notion of edges.

The location of affixes provides another example of the importance of edges in language. Across languages, prefixes and suffixes are much more frequent than infixes (Greenberg, 1957). If the representation of words use a positional memory mechanism as described above, one would expect affixes to be added in positions learners can better track, such as edges. By a similar account, one can imagine psychological accounts for other prominent linguistic phenomena, both phonological and, morphological. It is also possible to explain how different linguistic hierarchies are coordinated (Nespor & Vogel, 1986; McCarthy & Prince, 1993; see Endress, Nespor, & Mehler, 2009, for a review). For example, the morphosyntactic and phonological hierarchies do not always match (e.g., morphemes are not always syllables, as the English plural [s]). However, at least one of the edges of the constituents of these hierarchies always match; in the case of the plural [s], for instance, the right edge of the morpheme always coincides with the right edge of a syllable.

In sum, positional and chaining memory are one basis to explain participants' abilities in acquiring artificial languages reviewed above. They can also provide a natural account for most available empirical data. Finally, they might also have important ramifications for how real languages are represented and processed. In

contrast, not only do the alternative, single mechanism theories fail to provide an account for the majority of these findings, but it is also totally unclear how it would relate to real phenomena in real languages.

## Conclusions

In this paper, we reviewed evidence for the existence specialized learning mechanisms from artificial language learning, focusing particularly on the experiments following Peña et al. (2002) and Endress and Bonatti (2007). We have evaluated two prominent hypotheses — one more theoretical, and the other computational — assuming that one single, all-purpose statistical mechanism can explain language acquisition. We have shown that these hypotheses fail to account for most of the available evidence. In contrast, a multiple-mechanism account, which just assumes two well-known and independently motivated memory mechanisms, can explain most results reported in that literature. Further, it seems that the same mechanisms might play an important role in the learning and representation of real languages.

We do not claim that there might not be a single mechanism theory of a different kind that might account for the data. Nor do we claim that all or even most linguistic phenomena can be explained based on our hypothesis. Rather, we view the two mechanisms discussed here as just two elements of a much larger cognitive toolbox comprising a multitude of mechanisms, each of which being important to some linguistic phenomena and irrelevant to others. We surmise that learners can deal with the complexities of real language acquisition only by using the appropriate tools of their rich and powerful computational toolbox.

## References


Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170-176. Retrieved from http://cdp.sagepub.com/content/21/3/170.abstract doi: 10.1177/0963721412436806

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

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

Bates, E., & Elman, J. L. (1996). Learning rediscovered. *Science*, *274*(5294), 1849-50.

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. *Psychol Sci*, *16*(8).

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2007, Oct). On consonants, vowels, chickens, and eggs. *Psychol Sci*, *18*(10), 924–925. Retrieved from http://dx.doi.org/10.1111/j.1467-9280.2007.02002.x doi: 10.1111/j.1467-9280.2007.02002.x

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

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

Caramazza, A., Chialant, D., Capasso, R., & Miceli, G. (2000, Jan). Separable processing of consonants and vowels. *Nature*, *403*(6768), 428–430. Retrieved from http://dx.doi.org/10.1038/35000206 doi: 10.1038/35000206

Carreiras, M., & Price, C. J. (2008, Jul). Brain activation for consonants and vowels. *Cereb Cortex*, *18*(7), 1727–1735. Retrieved from http://dx.doi.org/10.1093/cercor/bhm202 doi: 10.1093/cercor/bhm202

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Conrad, R. (1960, Feb). Serial order intrusions in immediate memory. *Br J Psychol*, *51*, 45-8.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004, Sep). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *J Exp Psychol Learn Mem Cogn*, *30*(5), 1119-30. Retrieved from http://dx.doi.org/10.1037/0278-7393.30.5.1119 doi: 10.1037/0278-7393.30.5.1119

de Diego-Balaguer, R., Fuentemilla, L., & Rodriguez-Fornells, A. (2011, Oct). Brain dynamics sustaining rapid rule extraction from speech. *J Cogn Neurosci*, *23*(10), 3105–3120. Retrieved from http://dx.doi.org/10.1162/jocn.2011.21636 doi: 10.1162/jocn.2011.21636

de Diego Balaguer, R., Toro, J. M., Rodriguez-Fornells, A., & Bachoud-Lévi, A.-C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE*, *2*(11), e1175. Retrieved from http://dx.doi.org/10.1371/journal.pone.0001175 doi: 10.1371/journal.pone.0001175

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

Elman, J. L. (1990). Finding structure in time. *Cognit Sci*, *14*(2), 179–211. Retrieved from citeseer.ist.psu.edu/elman90finding.html

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

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

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A. D., Cahill, D., Block, S., Watumull, J., & Hauser, M. D. (2009, Dec). Evidence of an evolutionary precursor to human language affixation in a nonhuman primate. *Biol Lett*, *5*(6), 749–751.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007, Dec). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

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

Endress, A. D., & Langus, A. (under revieww). Transitional probabilities count more than frequency, but might not be used to learn words.

Endress, A. D., & Mehler, J. (2009a, Nov). Primitive computations in speech processing. *Q J Exp Psychol*, *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. *J Mem Lang*, *60*(3), 351-367.

Endress, A. D., Nespor, M., & Mehler, J. (2009, Aug). Perceptual and memory constraints on language acquisition. *Trends Cogn Sci*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *J Exp Psychol Gen*, *134*(3), 406-19.

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

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

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Goedemans, R., & van der Hulst, H. (2008a). Fixed stress locations. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online.* (chap. 14). Munich, Germany: Max Planck Digital Library. Retrieved from http://wals.info/feature/14 (Accessed on 2010-12-28)

Goedemans, R., & van der Hulst, H. (2008b). Weight-sensitive stress. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online.* (chap. 15). Munich, Germany: Max Planck Digital Library. Retrieved from http://wals.info/feature/15 (Accessed on 2010-12-28)

Greenberg, J. (1957). *Essays in linguistics.* Chicago: University of Chicago Press.

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.

Havy, M., Bertoncini, J., & Nazzi, T. (2011). Word learning and phonetic processing in preschool-age children. *Journal of Experimental Child Psychology*, *108*(1), 25 - 43. Retrieved from http://www.sciencedirect.com/science/article/B6WJ9-512DT3X-1/2/d95a0163753f176354d01ac537f931cf doi: DOI: 10.1016/j.jecp.2010.08.002

Hayes, B. (1995). *Metrical stress theory: Principles and case studies.* Chicago: University of Chicago Press.

Henson, R. (1998, Jul). Short-term memory for serial order: The Start-End Model. *Cognit Psychol*, *36*(2), 73-137.

Henson, R. (1999, Sep). Positional information in short-term memory: Relative or absolute? *Mem Cognit*, *27*(5), 915-27.

Hicks, R., Hakes, D., & Young, R. (1966, Jun). Generalization of serial position in rote serial learning. *J Exp Psychol*, *71*(6), 916-7.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011, Nov). Consonants and vowels: different roles in early language acquisition. *Dev Sci*, *14*(6), 1445–1458. Retrieved from http://dx.doi.org/10.1111/j.1467-7687.2011.01089.x doi: 10.1111/j.1467-7687.2011.01089.x

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007, Oct). Does grammar constrain statistical learning? Commentary on Bonatti, Peña, Nespor, and Mehler (2005). *Psychol Sci*, *18*(10), 922-3. Retrieved from http://dx.doi.org/10.1111/j.1467-9280.2007.02001.x doi: 10.1111/j.1467-9280.2007.02001.x

Laakso, A., & Calvo, P. (2011). How many mechanisms are needed to analyze speech? a connectionist simulation of structural rule learning in artificial language acquisition. *Cogn Sci*, *35*(7), 1243–1281. Retrieved from http://dx.doi.org/10.1111/j.1551-6709.2011.01191.x doi: 10.1111/j.1551-6709.2011.01191.x

Lenneberg, E. H. (1967). *Biological foundations of language.* New York: John Wiley and Sons.

Marchetto, E., & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, *67*(3), 130 - 150. Retrieved from http://www.sciencedirect.com/science/article/pii/S00100285130 doi: http://dx.doi.org/10.1016/j.cogpsych.2013.08.001

Marchetto, E., & Bonatti, L. L. (2015, Jul). Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities. *J Child Lang*, *42*(4), 873–902. Retrieved from http://dx.doi.org/10.1017/S0305000914000452 doi: 10.1017/S0305000914000452

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognit Psychol*, *37*(3), 243–82.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.

McCarthy, J. J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1993* (pp. 79–153). Boston, MA: Kluwer.

McClelland, J. L., & Patterson, K. (2002, Nov). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends Cogn Sci*, *6*(11), 465-472.

Mehler, J., & Dupoux, E. (1990). *Naître humain.* Paris: Odile Jacob.

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2008). The role of pause cues in language learning: The emergence of event-related potentials related to sequence processing. *Journal of Cognitive Neuroscience*, *20*(5), 892-905. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/jocn.2008.205 doi: 10.1162/jocn.2008.20511

Nazzi, T. (2005, Nov). Use of phonetic specificity during the acquisition of new words: differences between consonants and vowels. *Cognition*, *98*(1), 13–30. Retrieved from http://www.sciencedirect.com/science/article/B6T24-4F65K7R-6/1/6ad39551a66ac4e44368d956c89c0568

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Foris: Dordrecht.

New, B., Araújo, V., & Nazzi, T. (2008, Dec). Differential processing of consonants and vowels in lexical access through reading. *Psychol Sci*, *19*(12), 1223–1227. Retrieved from http://dx.doi.org/10.1111/j.1467-9280.2008.02228.x doi: 10.1111/j.1467-9280.2008.02228.x

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognit Psychol*, *48*(2), 127-62.

Ng, H. L., & Maybery, M. T. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology: Section A*, *55*(2), 391–424.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in speech processing. *J Mem Lang*, *53*(2), 225–237.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009, Nov). Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-7. Retrieved from http://dx.doi.org/10.1016/j.cognition.2009.07.011 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, Oct). A role for backward transitional probabilities in word segmentation? *Mem Cognit*, *36*(7), 1299–1305. Retrieved from http://dx.doi.org/10.3758/MC.36.7.1299 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. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X12000216 doi: http://dx.doi.org/10.1016/j.jml.2012.02.010

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004, Dec). Learning nonadjacent dependencies: No need for algebraic-like computations. *J Exp Psychol Gen*, *133*(4), 573-83. Retrieved from http://dx.doi.org/10.1037/0096-3445.133.4.573 doi: 10.1037/0096-3445.133.4.573

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *J Mem Lang*, *39*, 246–63.

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

Pinker, S. (1991). Rules of language. *Science*, *253*(5019), 530-5.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends Cogn Sci*, *6*(11), 456-463.

Pons, F., & Toro, J. M. (2010, Sep). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, *116*(3), 361–367. Retrieved from http://dx.doi.org/10.1016/j.cognition.2010.05.013 doi: 10.1016/j.cognition.2010.05.013

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *J Mem Lang*, *44*(4), 493–515.

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, Jan). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Dev Psychol*, *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. *J Mem Lang*, *35*, 606–21.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273 - 284.

Schulz, R. W. (1955, Apr). Generalization of serial position in rote serial learning. *J Exp Psychol*, *49*(4), 267-72.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599-603.

Seidenberg, M. S., & Elman, J. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*(5413), 433.

Shukla, M. (2006). *Prosodic constraints on statistical strategies in segmenting fluent speech* (Unpublished doctoral dissertation). International School for Advanced Studies, Trieste, Italy.

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

Swingley, D. (2005, Feb). Statistical clustering and the contents of the infant vocabulary. *Cognit Psychol*, *50*(1), 86-132. Retrieved from http://dx.doi.org/10.1016/j.cogpsych.2004.06.001 doi: 10.1016/j.cogpsych.2004.06.001

Toro, J. M., Bonatti, L., Nespor, M., & Mehler, J. (2008). Finding words and rules in a speech stream: functional differences between vowels and consonants. *Psychol Sci*, *19*, 137–144.

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008, Nov). The quest for generalizations over consonants: asymmetries between consonants and vowels are not the by-product of acoustic differences. *Percept Psychophys*, *70*(8), 1515–1525. Retrieved from http://dx.doi.org/10.3758/PP.70.8.1515 doi: 10.3758/PP.70.8.1515

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

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *J Exp Psychol: Hum Perc Perf*, *35*(1), 195–202.

Yang, C. D. (2004, Oct). Universal Grammar, statistics or both? *Trends Cogn Sci*, *8*(10), 451-456. Retrieved from http://dx.doi.org/10.1016/j.tics.2004.08.006 doi: 10.1016/j.tics.2004.08.006

Yang, C. D. (2010). Three factors in language variation. *Lingua*, *120*(5), 1160 - 1177. Retrieved from

http://www.sciencedirect.com/science/article/pii/S0024384109001120
doi: 10.1016/j.lingua.2008.09.015