Single vs. multiple mechanism models of artificial grammar learning

Ansgar D. Endress
Universitat Pompeu Fabra, Barcelona, Spain
City University, London, UK

Luca L. Bonatti
Universitat Pompeu Fabra, Barcelona, Spain

Draft of May 29, 2013

The extent to which language acquisition relies on computations that go beyond basic statistical abilities has been a topic of important debates. One crucial test case for such theories have been artificial language learning experiments. Recent computational models of artificial language learning suggested that not only the acquisition of words, but even of certain grammar-like regularities can be learned by basic statistical mechanisms. Here, we expand the scope of these models, asking whether they are compatible with human behavior. We show that even the results that were supposedly explained by statistical mechanisms contradict the modeling results. Through a more detailed analysis of previous simulations as well as new simulations, we demonstrate that these networks do not only fail to reproduce the data, but make predictions that are inconsistent with basic psychological facts. We suggest that the artificial language learning literature is better explained by multiple mechanism models, and that some of these mechanisms might draw on basic perceptual abilities that cannot be reduced to statistical computations.

Introduction

Language acquisition is a complex learning problem, and the underlying mechanisms are poorly understood. A useful theoretical baseline against which models of language acquisition can be compared are basic statistical processes, due to the widespread assumption that statistical processes are in some sense basic and domain-general (but see Garcia, Hankins, & Rusiniak, 1974; Gallistel, 2000; Gallistel & Gibbon, 2000). As such, before postulating complex and potentially language- and human-specific computational systems, it is often useful to ask whether non-statistical mechanisms are really required for rule-like regularities, or whether basic statistical computations might suffice.

The most prominent test case has been the mechanisms underlying the regular English past-tense inflections and other phenomena from inflectional morphology. It seems fair to say that the debate is still ongoing, and that a lot has been learned in the process (see McClelland & Patterson, 2002; Pinker & Ullman, 2002for reviews). Another research tradition that has been widely used to assess the need for non-statistical computations uses artificial languages. Such languages are highly simplified but mirror key characteristics of natural languages, and, importantly allow investigators to monitor the processes of language acquisition in the laboratory.

One such case study has been provided by Peña, Bonatti, Nespor, and Mehler (2002) and Endress and Bonatti (2007). They used the artificial language learning approach to study two crucial aspects of language acquisition: learning the words and the rules of a language. They concluded that independent mechanisms might underlie these different aspects of language acquisition. This research has led to a number of important challenges of different aspects of their claims. However, these criticisms have been addressed one by one, and do not seem to explain the available data (see below and Laakso & Calvo, 2011 for discussion).

A more important challenge has been put forward by Laakso and Calvo (2011). They replicated a subset of the simulations reported by Endress and Bonatti (2007), and, while finding similar results, concluded that one single mechanism could account for both aspects of language acquisition. However, the scope of their modeling was extremely limited, and did not take advantage of the extensive amount of empirical data available to assess the contribution of statistical and non-statistical mechanisms to language acquisition. Here, we build on this work, in an attempt to provide a more comprehensive test of the relative merits of single vs. multiple mechanism models of artificial grammar learning. We compare the simulations to the empirical data, derive

The research was supported by the Ministerio de Ciencia e Innovación Grant PSI2012-31961 to L.L.B and 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.

novel predictions from these simulations, and perform novel simulations.

Before discussing these results in more detail, it is useful to clarify the goals for the present work. Modelers of cognitive phenomena often propose to provide "existence proofs" showing that a single-mechanism, statistical learning model can account for data (e.g., Laakso & Calvo, 2011). However, as discussed extensively in, among others, Endress and Bonatti (2007), it is obvious that some network will reproduce some phenomenon or the other, and in fact, even Endress and Bonatti (2007) reported some simulations where the purely statistical models partially reproduced some aspects of the data they used to argue for a multiple mechanism theory. Here, we assume that, for a model to have any scientific interest at all, it is not sufficient to partially reproduce one phenomenon out of a multitude of experiments, but that statistical models need to be compared against a wider set of data than previously accomplished.

Learning words and rules from fluent speech

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 (Aslin, Saffran, & Newport, 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; but see Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Pilon, 1981). Moreover, while linguists had noted that distributional analysis might lead to cues to word boundaries (e.g., Harris, 1955), such observation were generally deemed psychologically implausible. This view changed with the seminal demonstration that even young infants are sensitive to distributional cues (Aslin et al., 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996). These results opened the possibility 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; Hauser, Newport, & Aslin, 2001; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran & Griepentrog, 2001; Toro & Trobalón, 2005) and, consequently, a widespread consensus that infants can use TPs to extract words from fluent speech (but see Endress & Mehler, 2009b; Medina, Snedeker, Trueswell, & Gleitman, 2011; Gillette, Gleitman, Gleitman, & Lederer, 1999; Yang, 2004, for opposing views).

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; Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Bonatti, Nespor, & Mehler, 2008; Toro, Shukla, Nespor, & Endress, 2008).

Extracting words and rules from artificial speech streams

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. Exposing human adults to a sequence of trisyllabic items, these authors first showed that human adults can find words in a continuous speech stream by tracking statistical dependencies between nonadjacent items. (Different authors challenged the conclusion that participants can track TPs between non-adjacent syllables, and suggested that the relevant data might rather be a result of phonological confounds (Newport & Aslin, 2004; Onnis, Monaghan, Richmond, & Chater, 2005; Perruchet, Tyler, Galland, & Peereman, 2004). However, Peña et al. (2002) had actually controlled for such confounds, and it is now fairly accepted even by scholars who were initially skeptical of such abilities that participants are indeed sensitive to such TPs (e.g., Peña et al., 2002; Endress & Bonatti, 2007; Endress & Mehler, 2009a; Endress & Wood, 2011; Onnis et al., 2005; Pacton & Perruchet, 2008), a conclusion that is shared even by scholars who are otherwise critical of the multiple mechanisms hypothesis (e.g., Laakso & Calvo, 2011). We thus do not discuss this literature further.)

Importantly, Peña et al. (2002) also showed that participants could not use the abilities to track TPs be-

tween nonadjacent items 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 learners to extract the hidden generalizations. 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 the type of statistical mechanisms that supports TP computations, but rather required different mechanisms. Endress and Bonatti (2007) dubbed this conclusion 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 additional markers, such as short silences, are inserted between words, a second (class of) mechanism(s) 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.

To test these issues, Peña et al. (2002) and Endress and Bonatti (2007) familiarized participants with a string of trisyllabic words. Words were constructed from three $A_i \dots C_i$ frames, into which three different syllables could be inserted. This yielded 9 words of the form $A_i X C_i$. During the test phase following the familiarization, Peña et al. (2002) and Endress and Bonatti (2007) used three critical types of test items among which participants had to choose: "class-words," "part-words" and "rule-words." Class-words had the structure $A_i X' C_j$. 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 $C_i A_j X$, taking the last syllable from one word and the first two syllables from the next word, or the structure $X C_i A_j$, 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. However, 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 $A_i X' C_i$; hence, they had "correct" initial and final syllables, and TPs of 1.0 between their first and their last syllable.

To test whether participants were 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 evidence 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 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 had occurred during the familiarization and had, therefore, non-zero TPs. Results showed that participants preferred class-words to part-words only when familiarized with a segmented stream, in which 25 ms silences were inserted between words, 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 their 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 $A_i X Y Z C_i$, 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 $X A_i Y C_i Z$). It should be noted that all of these results have ben replicated with non-linguistic stimuli such as action sequences (Endress & Wood, 2011).

Strikingly, 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

Table 1
Summary of the main test item types used by Peña et al. (2002) and Endress & Bonatti (2007).

Words	Part-Words	Rule-Words	Class-Words
$\mathbf{A}_i\mathrm{X}\mathbf{C}_i$	$\mathbf{C}_i\|\mathbf{A}_j\mathrm{X}$ or $\mathrm{X}\mathbf{C}_i\|\mathbf{A}_j$	$\mathbf{A}_i\mathrm{X}'\mathbf{C}_i$	$\mathbf{A}_i\mathrm{X}'\mathbf{C}_j$
Appear in the stream $\mathrm{TP}(\mathrm{A}_i \rightarrow \mathrm{C}_i) = 1$	Appear in the stream but straddle a word boundary	As words, but with new X syllables	As rule-words, but with first and last syllable from different families

available very quickly, whereas statistical information appears to require time and exposure to consolidate.[1]

Prior simulations of rule acquisition with a single statistical mechanism

While such data suggest several mechanisms might be at play in artificial language learning (and, by extension, in language acquisition in general), they do not provide a definitive proof. Endress and Bonatti (2007) provided other indirect evidence that a single statistical mechanism could not account for these data, based on simulations with a prominent candidate single-mechanism model that has been widely used in the study of language acquisition: a Simple Recurrent Network (SRN; Elman, 1990). In line with previous research, the network was 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, using 20 simulations as units of analysis in order to match the statistical power of their behavioral experiments. The results clearly showed that the overall pattern of the simulations was not compatible with the behavioral data.

Importantly, Endress and Bonatti (2007) did not use these results to argue that SRN cannot prefer class-words to part-words in principle. In fact, they found conditions in which the network did mimic certain aspects of the participants' responses. However, they also showed that the conditions under which such aspects of the data were successfully reproduced were inconsistent with other aspect of the data. The networks seemed to reproduce the preference for class-words over part-words under specific assumptions:

1. The network had to encode silences in the familiarization with an explicit extra-symbol.

2. The test part-word did not include silences (i.e., the extra-symbol).

3. Class-words were compared only against part-words of type XC_iA_j but not to part-words of type C_iA_jX.

All three conditions are problematic. First, while the silences clearly influenced the computations performed by the participants, several considerations suggest that they are not represented as an extra-symbol that undergoes TP computations just like syllables; we will review this evidence below. Second, while the participants' performance did not depend on whether or not the part-words included silences (see Peña et al.'s (2002) Footnote 27 and below), the network behaved in a markedly different way depending on whether the silences were included in the test items. Third, in contrast to the network, adults' preference for class-words over part-words never depended on the type of part-word against which class-words were tested. For these reasons, Endress and Bonatti (2007) concluded that the results of the simulation "suggest that purely statistical mechanisms such as SRNs cannot account for the preference for class-words or for the negative correlation between the preference for class-words and the familiarization duration" (p. 285). Again, this conclusion was not a principled argument against the possibility that a single mechanism could account for some aspect of the data or the other, but an overall assessment of the behavior of a large class of networks under different simulation constraints, once the available evidence was taken into account.

Laakso and Calvo (2011) recently revisited the question of whether an SRN could account for the overall data. They ran a set of simulations that was very similar

[1] These results generalizes 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, while participants preferred rule-words to part-words only when exposed to a brief segmented stream, they 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.

to those reported by Endress and Bonatti (2007). They studied a slightly different network, in which the number of hidden units, the activation function as well as the number of simulations taken as units for analysis were changed with respect to Endress and Bonatti (2007). Aside from these differences, Laakso and Calvo (2011) basically replicated a subset of Endress and Bonatti's (2007) simulations.

As expected, their results were similar to Endress and Bonatti's (2007). However, they drew markedly different conclusions. Specifically, according to Laakso and Calvo (2011), their network allegedly preferred class-word to part-words after few training cycles but inverted this preference after more familiarization cycles, ostensibly reproducing one aspect of Endress and Bonatti's (2007) results with adult participants. However, closer examination of the modeling results revealed that the network's success in reproducing this aspect of the results was only partial, because the reversal in the preference was carried *exclusively* by part-words of type C_iA_jX, whereas part-words of type XC_iA_j were never preferred to class-words after any number of training cycles. Nevertheless, because the reversal in the preference, albeit for both part-word types, was one of the arguments Endress and Bonatti (2007) marshaled in support of the presence of multiple mechanisms, Laakso and Calvo (2011) took their network's partial success as a proof that language can be acquired by means of one single statistical mechanism.

Given the limited scope of Laakso and Calvo's (2011) simulations, we will ask whether such models account for the qualitative pattern of results in Endress and Bonatti's (2007) and other studies, and whether these data are really consistent with Laakso and Calvo's (2011) simulations, broadening the scope of the data to which the network results are compared.

Laakso and Calvo's (2011) simulations

We first analyze those of Endress and Bonatti's (2007) experiments that Laakso and Calvo (2011) simulated, asking whether the experiments are really consistent with the simulations.

Asymmetries between test items types. 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, allegedly reproducing some of Endress and Bonatti's (2007) data. This claim, however, is incorrect. As is clear from their Figure 4, and as discussed by Endress and Bonatti (2007) (p. 283), the network reversed the preference only against part-words with the structure C_iA_jX, while the network preferred 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.[2]

Laakso and Calvo (2011) tried to explain away the apparent discrepancy between the data and their central result arguing that Endress and Bonatti's (2007) data were not so compelling after all, and 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, it is clearly ad-hoc, and Laakso and Calvo (2011) did not provide any evidence to support it. Further, Laakso and Calvo (2011) did not perform a power analysis to evaluate the hypothesis that the relevant experiments lack statistical power; they did not specify what level of statistical power would be sufficient for an adequate test of the hypotheses or the model; and they did not show that the test they proposed addresses the alleged problem of statistical power in any way. Thus, it appears that the claim that Endress and Bonatti's (2007) data lack adequate statistical power is not motivated, and the actual results seem to show that Laakso and Calvo's (2011) model contradicts the available evidence.

The dynamics of the preference for different test items. While Laakso and Calvo's (2011) arguments are mainly based on the dynamics of their networks, this dynamics reveals qualitative departures from the empirical results. For example, in their Study 2, their network was exposed to a continuous speech stream and part-words were preferred to class-words after the earliest familiarization durations. In contrast, in Endress and Bonatti's (2007) results, participants remained at chance even after a 10 min familiarization. Likewise, in their Study 2, part-words are preferred to rule-words after few familiarization cycles. This contrasts markedly with Peña et al.'s (2002) results where such a preference arose only after very long familiarization durations of at least 30 min. Hence, the network behavior is inconsistent with human behavior even for the network dynamics which Laakso and Calvo (2011) present as their strongest case.

Laakso and Calvo (2011) acknowledge the latter discrepancy between the model and the data, and speculate about the psychological processes of Endress and Bon-

[2] 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).

atti's (2007) participants. They argue that "based on [their] network modeling results, however, [they] have suggested that participants may not be learning a class rule even when the familiarization stream contains segmentation cues. Rather, participants may simply be slower to develop a dispreference for class words than they are to develop a dispreference for part words." (p. 23). It appears curious to make inferences about what humans learn "based on network modeling results" that do not fit the humans' behavior in the first place. It thus seems fair to conclude that their model is at odds with the available empirical data.

The relative preferences for different test items. It is also instructive to ask whether the network correctly predicts the relative strength of the preferences for different test items. 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. To compare the strength of the model's preferences with those of actual humans, we compared the effect sizes of these results. 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. Actual humans show 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. (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 hence does not show the marked advantage for the class-word vs. part-word discrimination shown by Laakso and Calvo's (2011) network.) Thus, the model behavior does not fit the human data.

Is it harder to remember words heard more often?. 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, consider that, in some of Endress and Mehler's (2009a) experiments, the familiarization stream consisted of a clearly distinguishable sequence of words presented in isolation, separated by silences of 1 s. Hence, the subsequent two-alternative forced-choice task just amounted to a memory test for words.

If, as Laakso and Calvo (2011) propose (and we agree) the 1-s separation is computationally equivalent to the 25-ms silences used by Endress and Bonatti (2007), their model ought to apply to Endress and Mehler's (2009a) experiments as it does to Endress and Bonatti's (2007) experiments. As a result, the prediction that the preference for words over part-words decreases with longer familiarizations just amounts to the prediction 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. Hence, the network behavior contradicts one of the best-established facts of experimental psychology.

Expanding the scope of the investigation: Further explorations of the SRN model

Peña et al. (2002), Endress and Bonatti (2007), Endress and Mehler (2009a) and Endress and Wood (2011) provided other data that have not been compared against statistical learning models. We will now turn to these results

For example, in their Footnote 27, Peña et al. (2002) 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 Laakso and Calvo's (2011) model. Given that the model 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. In fact, Endress and Bonatti (2007) investigated this issue in their simulations. As is clear from their Figure 11, the network settings in which the network prefers class-words to part-words essentially disappear once silences are included in the part-words, providing another case for the MOM hypothesis.

Another critical case for the MOM hypothesis comes from Endress and Mehler's (2009a) data. They used penta-syllabic words (as opposed to the trisyllabic items used by Endress & Bonatti, 2007). That is, while Endress and Bonatti (2007) used words of the form A_iXC_i, Endress and Mehler (2009a) used words of the form A_iXYZC_i (where the critical A and C syllable were in the first and the last position), and XA_iYC_iZ (where the critical syllables were in word-internal positions). Endress and Mehler (2009a) 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 seg-

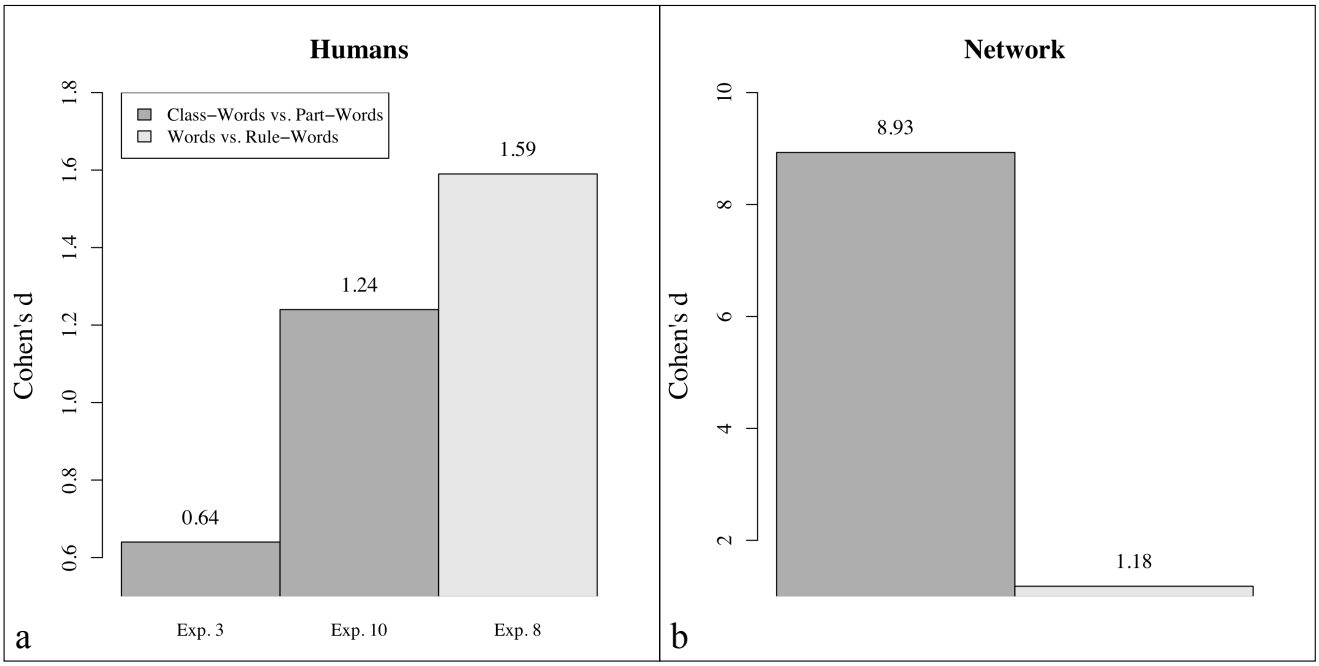

Figure 1. Effect sizes (Cohen's d) for the class-word vs. part-word discrimination (dark bars) and the word vs. rule-word discrimination with segmented 2-min streams in (a) humans and (b) networks. (a) In humans, the word vs. rule-word discrimination is numerically easier than the class-words vs. part-word 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.

mented stream, participants preferred class-words to part-words, but only when the critical syllables were in the first and the last 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.

Laakso and Calvo (2011) claim that "the results of Endress and Mehler can easily be accommodated within the general framework herewith advocated" (p. 30). However, this is most likely false. 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 to 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). Laakso and Calvo's (2011) 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.[3]

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 of the simulations, the preference for class-words was at least numerically stronger in the middle condition than in the edge condition (see Figure 3). [4] Hence, it seems that an SRN does not easily account for

[3] 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.

[4] 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 sim-

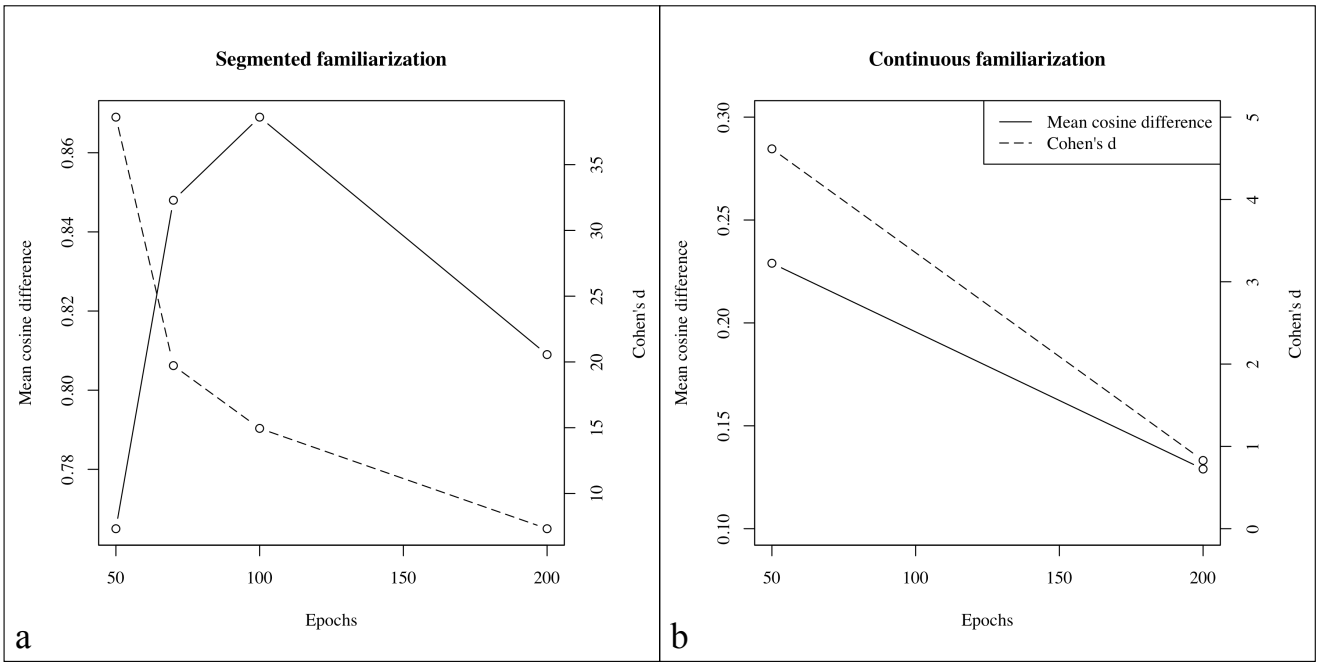

Figure 2. Difference in predictions for the last syllable of words and part-words, respectively, after (a) a familiarization with a segmented stream and (b) a familiarization with a continuous stream. The solid line shows the average difference between the cosinus values between the predicted network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

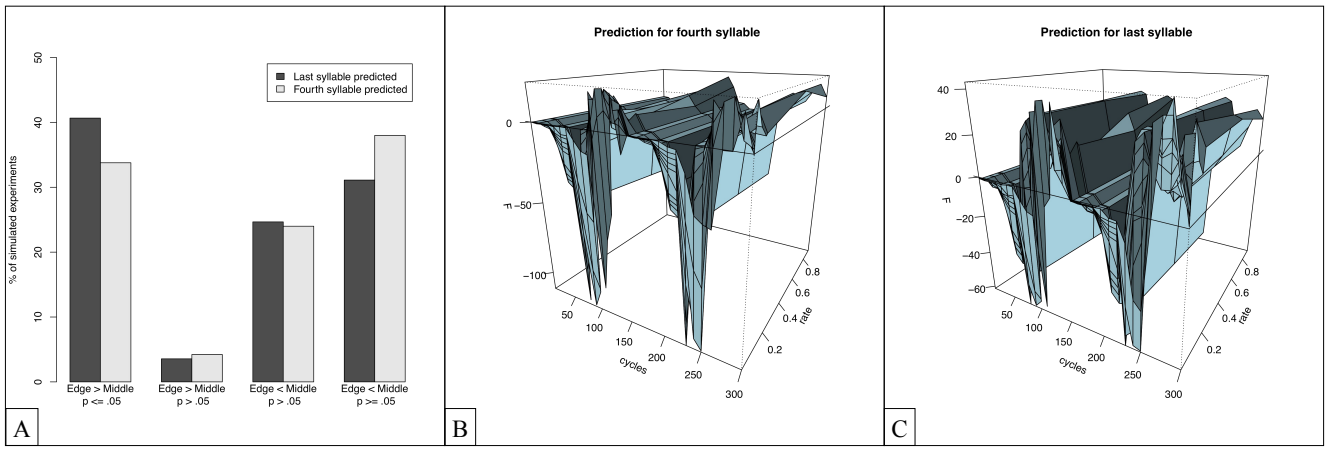

Figure 3. We simulated experiments by recording the results of 20 simulations with different network initializations, representing 20 participants. One experiment was simulated for each set of network parameters. The networks did not reproduce the preference for class-words over part-words in the edge 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.

Endress and Mehler's (2009a) results. In other words, the network fails to reproduce the basic psychological phenomenon that events in edges of sequences are easier to process than edge in sequence-middles.[5]

The data that every theory should model: the converging evidence for multiple learning mechanisms

The analyses so far focused on phenomena that might potentially be modeled by an SRN. However, there are other strands of research that support the multiple mechanism view for which it is not even clear how they can possibly modeled by such a model.

For example, the two mechanisms seem to have a different developmental time course. When 18-months-old infants are exposed to artificial streams similar to Peña et al.'s (2002) stimuli, but containing a conflict between statistical information and generalizations, 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. In contrast, 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, under review).[6] 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.

Likewise, 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 tested on *CBA vs. FED*.[7] 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.

Further, 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.

Another piece of evidence for independent mechanisms comes from brain imaging experiments. For example, using material similar to Peña et al.'s (2002), different authors suggested that ERPs differ according to whether participants extract words or rules from the same speech stream (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

ilarity 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 completeness, we represent both possibilities in Figure 3

[5] 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.

[6] 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 & Bonatti's experiments relied on statistical relations among non-adjacent syllables, and such relations are likely to be more difficult to track.

[7] 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.

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.

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 (see Table 2 for a summary). These dissociations appear in 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, and the specificity of representations they create.

Further, even among the aspects of the data that have been modeled with an SRN, the model behavior differs qualitatively from the behavior of actual humans, and makes predictions that are at odds with basic psychological phenomena, such as the prediction that memory for items should be worse when they are repeated more often, or the failure to reproduce the advantage for processing items in edges of sequences. Hence, the model does not only clash with some opaque details of the available data, but even the most crucial arguments put forward in support of a single mechanism account seem to qualitatively contradict the available data.

In contrast, there might be a simple explanation of the facts reviewed above, based on two well-known types of memory encoding for sequences, 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. While this account does not *predict* all of the data reviewed above (e.g., there is no a priori reason to predict that the two mechanisms are differentially sensitive to spatial rotation), it is at least consistent with it.

More generally, we believe that the issue of how many mechanisms exist and how they work together cannot be decided by "existence proofs" alone, if such existence proofs are constrained to an extremely limited aspect of the available data. Rather, theories need to be tested against the available data, and should be compatible with basic psychological facts. This does not show that purely associationnist accounts of these data are necessarily incorrect, but it does suggest that artificially restricting the debate to limited sets of partial simulations is a non-starter.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170-176.

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.

Balaguer, R. D. D., 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.

Batchelder, E. O.(2002). 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. *Psychological Science*, *16*(8).

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. *Language and Speech*, *54*(1), 49–72.

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

Conrad, R.(1960). Serial order intrusions in immediate memory. *British Journal of Psychology*, *51*, 45-8.

de Diego-Balaguer, R., Fuentemilla, L., & Rodriguez-Fornells, A. (2011). Brain dynamics sustaining rapid rule extraction from speech. *Journal of Cognitive Neuroscience*, *23*(10), 3105–3120.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

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

Table 2
Summary of dissociations between the statistical mechanism(s) and the mechanism(s) responsible for the generalizations.

	Statistical mechanism	**Generalization mechanism(s)**	**References[a]**
Requires explicit boundary cues	–	+	PNMB, EB, EM, EW
Available after brief exposure	(–)	+	PNMB, EB
Operates in sequences edges	+	+	EM, EW
Operates in sequences edges	+	–	EM
Tolerates spatial rotation	–	+	EW
Tolerates temporal reversal	+	–	EW
Present at 18 months	+	+	MB
Present at 12 months	–	+	MB
Has different brain correlates	+	+	DFR, MBF

[a]PNBM: Peña et al. (2002); EB: Endress and Bonatti (2007); EM: Endress and Mehler (2009a); EW: Endress and Wood (2011); DFR: Balaguer et al. (2007); MBF: Mueller et al. (2008)

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

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., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

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

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

Gallistel, C. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (2nd ed., pp. 1179–91). Cambridge, MA: MIT Press.

Gallistel, C., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289-344.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1974). Behavioral regulation of the milieu interne in man and rat. *Science*, *185*(4154), 824-31.

Gillette, J., Gleitman, H., Gleitman, L. R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-76.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*, 190–222.

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.

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

Henson, R. (1999). Positional information in short-term memory: Relative or absolute? *Memory and Cognition*, *27*(5), 915-27.

Hicks, R., Hakes, D., & Young, R. (1966). Generalization of serial position in rote serial learning. *Journal of Experimental Psychology*, *71*(6), 916-7.

Laakso, A., & Calvo, P. (2011). How many mechanisms are needed to analyze speech? a connectionist simulation of structural rule learning in artificial language acquisition. *Cognitive Science*, *35*(7), 1243–1281.

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

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *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.

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

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9014–9019.

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.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *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. *Journal of Memory and Language*, *53*(2), 225–237.

Pacton, S., & Perruchet, P. (2008). An attention-based asso-

ciative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(1), 80-96.

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

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology. General*, *133*(4), 573-83.

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.

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 in Cognitive Sciences*, *6*(11), 456-463.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *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). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74-85.

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.

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). Generalization of serial position in rote serial learning. *Journal of Experimental Psychology*, *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.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86-132.

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

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008). The quest for generalizations over consonants: asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception and Psychophysics*, *70*(8), 1515–1525.

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., & 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.

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