COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 1243–1281

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01191.x

How Many Mechanisms Are Needed to Analyze Speech? A Connectionist Simulation of Structural Rule Learning in Artificial Language Acquisition

Aarre Laakso,[a] Paco Calvo[b]

[a]*Department of Behavioral Sciences, University of Michigan-Dearborn*
[b]*Philosophy Department, University of Murcia*

Received 29 April 2010; received in revised form 28 January 2011; accepted 18 April 2011

Abstract

Some empirical evidence in the artificial language acquisition literature has been taken to suggest that statistical learning mechanisms are insufficient for extracting structural information from an artificial language. According to the more than one mechanism (MOM) hypothesis, at least two mechanisms are required in order to acquire language from speech: (a) a statistical mechanism for speech segmentation; and (b) an additional rule-following mechanism in order to induce grammatical regularities. In this article, we present a set of neural network studies demonstrating that a single statistical mechanism can mimic the *apparent* discovery of structural regularities, beyond the segmentation of speech. We argue that our results undermine one argument for the MOM hypothesis.

Keywords: Artificial grammar learning; Speech processing; Language acquisition; More than one mechanism hypothesis; Statistical learning; Connectionism

1. Introduction

Over the last 20 years, the great debate about the architecture of cognition (e.g., Fodor & Pylyshyn, 1988; McClelland & Rumelhart, 1986) has remained at the forefront of work on language acquisition and speech processing. The question is whether speech can be processed—or indeed whether language could be acquired—solely by statistical learning mechanisms, or whether rule-following mechanisms are required. Evidence that both infants

Correspondence should be send to Aarre Laakso, Department of Behavioral Sciences, University of Michigan-Dearborn, 4020 CASL Building, 4901 Evergreen Road, Dearborn, MI 48128. E-mail: aarre@umich.edu

and adults can segment speech using statistical mechanisms like the computation of transitional probabilities (TPs) among syllables (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) convinced many that statistical mechanisms are involved in language acquisition and speech processing. Numerous papers in the last two decades demonstrate the power of statistical mechanisms (to name but a few: Christiansen & Chater, 1999; Hare, Elman, & Daugherty, 1995; Plunkett & Juola, 1999; Redington, Chater, & Finch, 1998; Seidenberg, 1997). In fact, in recent years, more sophisticated forms of statistical learning have been reported. Both infants and adults can even segment speech using backward, rather than forward, TPs (Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008). The debate then focused on whether statistical mechanisms are sufficient for the purposes of language acquisition and speech processing. The exchange about ''rule learning'' in infants (Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Seidenberg & Elman, 1999a) is one illustration.

Notwithstanding the consensus that statistical mechanisms could lie behind word segmentation (see below), the claim that a statistical mechanism might suffice for the induction of grammar remains contentious. Poverty-of-the-stimulus arguments (e.g., Chomsky, 1980) have convinced many that purely statistical mechanisms embedded in artificial neural networks (Seidenberg & Elman, 1999a, 1999b) or in human subjects (Saffran, Newport, et al., 1996) cannot account for the acquisition of language *as a whole*. Although speech may be segmented on the basis of distributional information, rule-following mechanisms that rest upon the manipulation of symbolic structures seem to be required in order to perform grammatical induction.

In this environment, Peña, Bonatti, Nespor, and Mehler (2002a), and more recently Endress and Bonatti (2007), are among those who have defended a third, ''hybrid'' possibility: that statistical and symbolic mechanisms might work together in a compound cognitive architecture. In their view, multiple mechanisms are needed in order to account for the data on adults' ability to acquire an artificial language. Peña et al. (2002a) report experimental evidence that they argue shows that humans use both statistical learning (to segment speech) and algebraic computations (to induce structural regularities like grammatical rules). That is, they argue that statistics are insufficient to support the discovery of underlying grammatical regularities, and that their results imply knowledge of rules. In particular, Peña et al. (2002a) designed a set of experiments aimed at assessing whether statistical computations based on TPs of the sort that are exploited in speech segmentation (Saffran, Aslin, et al., 1996) could also be used in order to induce rule-like regularities in the speech stream. They familiarized participants with a continuous sequence of artificial ''words,'' where what counts as a ''word'' is a function of the TPs between specific non-adjacent items (see Section 1.1 below for the details). In the test phase, participants were asked to choose, from between pairs of stimuli, which seemed more like a word from the familiarization stream. For example, one condition required participants to choose between words and items that had appeared in the familiarization stream but had straddled a word boundary. Another required participants to choose between items that had appeared in the familiarization stream but straddled a word boundary and items that had *not* appeared in the familiarization stream but respected the TP between specific non-adjacent items, as in the case of words.

The objective was to test subjects' ability to segment speech and to generalize beyond the familiarization stream.

Although participants were able to extract the lexicon based on non-adjacent dependencies (they chose words when compared to other familiarization items that straddled a word boundary), they failed to generalize beyond the familiarization corpus (did not choose items that had *not* appeared in the familiarization stream, but that respected the TP between specific non-adjacent items when compared to familiarization items that straddled a word boundary). Participants were, however, able to generalize when subliminal pauses indicating word segmentation boundaries were inserted into the corpus. Furthermore, participants were unable to induce the rule even when the duration of familiarization on a continuous stream was increased substantially. By contrast, participants could generalize after much shorter durations when the familiarization stream was segmented using subliminal cues. Peña et al., therefore, claim that statistics are not sufficient to extract structural information from a continuous familiarization corpus.

Endress and Bonatti (2007) replicate and extend the Peña et al. results, and attempt to model them using connectionist networks, taking the argument a step further by claiming that participants may be tuning to rules at an even higher level of abstraction than Peña et al. had proposed. Endress and Bonatti are unable to model the experimental results with connectionist networks, and they claim that their failure demonstrates that associative learning mechanisms are insufficient for language learning. They advocate instead what they call the ''more than one mechanism'' (MOM) hypothesis, according to which two different computational mechanisms must be responsible for the results they report: (a) a statistical mechanism for performing speech segmentation; and (b) a rule-governed mechanism responsible for the induction of grammatical or structural regularities in speech.

We found it surprising that a connectionist network was unable to model human performance in this task. We also observed that, although Endress and Bonatti had used a range of network parameters in their simulations, it would be impossible to test all possible values of every possible network parameter. In particular, we observed that Endress and Bonatti had used a relatively small number of hidden units (at most 27) to model the task, and that they had used only one of many possible combinations of activation function and error function. We therefore set out to determine whether, by manipulating network parameters, including the number of hidden units and the activation and error functions, we could model the behavioral data using a connectionist network.

Indeed, in this article, we report a set of connectionist simulations—based on a *single* statistical mechanism—that *does* model the experimental results of Peña et al. and Endress and Bonatti. We conclude that they have not demonstrated that rule-governed structure learning mechanisms are necessary for artificial language acquisition, as their MOM hypothesis suggests. The structure of this paper is as follows. In the remainder of the Introduction, we provide an overview of the evidence for the MOM hypothesis and explain the challenge for statistical learning. In Section 2, we report a preliminary analysis of the types of dependencies in the familiarization corpus in the relevant experiments. In the third section, we report the results of two simulations using artificial neural networks. In the general discussion, we argue that our results, insofar as the connectionist model employed does not implement

rules, undermine Peña et al. and Endress and Bonatti's argument for the MOM hypothesis. We argue, moreover, that their evidence is mostly negative (cases of generalization where a statistical explanation is not immediately forthcoming), and that when an attempt is made to provide positive evidence for the MOM hypothesis, it can easily be accommodated within the more parsimonious general framework we advocate here. Directions for future research and conclusions follow.

1.1. An overview of the evidence for the MOM hypothesis

Peña et al. (2002a) tested adults' abilities to segment speech based on non-adjacent dependencies and to generalize beyond the familiarization corpus. The experiments were based on roughly the same method as that used by Newport and Aslin (2000). Adult participants were asked to listen to a sequence of trisyllabic artificial ''words'' for a certain period of time. The artificial ''words'' had the form A_iXC_i, where A_i, X, and C_i are syllables. The subscripts on A and C indicate that the non-adjacent syllables are matched, such that the TP between an A_i and the following C_i is 1.0. There are three X syllables, so the TPs between an A_i and an intermediate X and between X and the final C_i are each 0.33. There are three word classes ($i \in [1,2,3]$), and no two adjacent words in the speech stream may be from the same class, so the TPs between the final syllable of one word C_i and the first syllable of the next word A_j is 0.5. The three word classes are *pu…ki*, *be…ga*, and *ta…du*. The three filler syllables are *li*, *ra*, and *fo*. Thus, the A_1XC_1 family consists of the words *puliki*, *puraki*, and *pufoki*; the A_2XC_2 family consists of the words *beliga*, *beraga*, and *befoga*; and the A_3XC_3 family consists of the words *talidu*, *taradu*, and *tafodu*. Familiarization streams were produced by concatenating speech-synthesized tokens of these nine words. After familiarization, participants were asked to choose, between pairs of stimuli, those that seemed more like a word from the familiarization stream. Test stimuli were of three kinds: ''words'' (items of the form A_iXC_i that had appeared in the familiarization stream), ''part words'' (items that had appeared in the familiarization stream but straddled a word boundary), and ''rule words'' (items of the form $A_iX'C_i$ that had *not* appeared in the familiarization stream, where X′ stands for a ''familiar'' syllable, that is, one that occurs in familiarization, although never between A_i and C_i).

In Peña et al.'s Experiment 1, participants were familiarized with a continuous speech stream, as described above, for 10 min. In the test phase, they were asked to choose between a ''word'' and a ''part word.'' The result was that participants preferred words over part words. Experiment 1 is consistent with the hypothesis that statistics alone suffice for speech segmentation, because what counts as a ''word'' is a function of the TPs between specific non-adjacent items (in this case, between A_i and C_i).

In their Experiment 2, Peña et al. investigated whether participants were simply segmenting the stream by exploiting different TPs between words and part words, or whether they were attuning to some more abstract underlying grammatical regularity. In order to answer this question, participants were asked, after having been familiarized for 10 min to the same speech stream as in Experiment 1, to choose between a part word and what Peña et al. dubbed a ''rule word.'' As it turned out, participants preferred part words over rule words,

suggesting that they had not extracted the rule from the training stream even though (as shown in Experiment 1) they had learned to segment the words of the language. Peña et al., therefore, claim that their Experiment 2 shows that statistics are not sufficient to extract structural information from a continuous familiarization corpus. In light of Experiments 1 and 2 together, Peña et al. conclude that a ''computational mechanism sufficiently powerful to support segmentation on the basis of nonadjacent TPs [experiment 1] is insufficient to support the discovery of the underlying grammatical-like regularity embedded in a continuous speech stream [experiment 2]'' (p. 605).

In Peña et al.'s Experiment 3, a ''subliminal'' 25 ms pause was inserted between each pair of words in the familiarization stream. Although participants reported no awareness of such gaps, their presence did affect the results. When participants were trained on a speech stream with gaps, the participants subsequently preferred rule words to part words at test. In their view, these results imply knowledge of rules, insofar as the very notion of an abstract rule word underlies the successful discrimination of rule words and part words. Thus, they write, ''This seems to be due to the fact that the selected items are compatible with a generalization of the kind 'If there is a [pu] now, then there will be a [ki] after an intervening X''' (p. 606). In other words, Peña et al. contend that two different computational mechanisms must be responsible for the results of Experiments 1–3: (a) a statistical mechanism for performing speech segmentation (Experiment 1); and (b) a rule-governed mechanism responsible for the induction of grammatical structural regularities in the corpus (Experiment 3).

Peña et al.'s Experiments 4 and 5 tested for preference between part words and rule words after familiarization on a continuous stream for 30 min or on a segmented stream for 2 min (see Table 1). In the first case, participants preferred part words over rule words, demonstrating that even lengthy familiarization with a corpus that does not contain prosodic cues to segmentation does not lead to abstraction of the rule. In the second case, participants preferred rule words over part words, demonstrating that even very brief familiarization with a corpus that does contain cues to segmentation leads to abstraction of the rule. Peña et al. interpret the results in Table 1 as evidence for the MOM hypothesis: a statistical mechanism for segmenting the familiarization corpus (Experiment 1), and a rule-governed mechanism that accounts for the induction of the rule that prefers rule words over part words (Experiments 3 and 5).

Furthermore, Endress and Bonatti argue that participants may not prefer rule words *themselves*, but so-called class words. Class words have the form $A_iX'C_j$, that is, an A syllable

Table 1
Summary of Peña et al.'s experimental results

Experiment	Stream	Familiarization Duration	Test Choice
1	Continuous	10 min	Words over part words
2	Continuous	10 min	No preference between rule words and part words
3	Segmented	10 min	Rule words over part words
4	Continuous	30 min	Part words over rule words
5	Segmented	2 min	Rule words over part words

from one class, followed by a syllable that had appeared in the speech stream but never in the middle of a word (as in rule words) from a different class, followed by a C syllable from the third class. These are called ''class words'' because they would be preferred if participants learned rules of the form ''if the first syllable is from the A class, then the last syllable is from the C class.''

The constraint that the second syllable must come from a different class than both the first syllable and the third syllable is not stated explicitly by Endress and Bonatti (2007), who define class words as ''items with the structure $A_iX'C_j$; A_i and C_j always occurred, respectively in the first and third positions of words in the stream but never in the same word, and X' is a syllable that occurred in the stream but never in the middle position of words'' (p. 251). Nevertheless, this constraint is implicit in the list of class word test items in their Appendix A. There would be 18 class words given the explicit definition that appears in Endress and Bonatti's (2007) paper: *beduki*, *bekidu*, *bepudu, bepuki*, *betadu*, *betaki*, *pubedu*, *pubega*, *puduga*, *pugadu*, *putadu*, *putaga*, *tabega*, *tabeki*, *tagaki*, *takiga*, *tapuga*, and *tapuki*. However, we learned via private correspondence that the set of class words that Endress and Bonatti actually used in their experiments incorporates the additional constraint that the second syllable cannot be from the same class as either the first syllable or the second syllable. This constraint reduces the number of class words to 12: *beduki*, *bekidu*, *bepudu*, *betaki*, *pubedu*, *puduga*, *pugadu*, *putaga*, *tabeki*, *tagaki*, *takiga*, and *tapuga*. We are not sure why the authors imposed this constraint, because the whole idea behind testing class words is that the middle syllable does not matter—that participants may ''have learned that the first and the last position in a word are variables that take their values from distinct classes'' (p. 251). Nevertheless, in the studies reported below, we describe tests using only the same restricted set of 12 class words that Endress and Bonatti used.

Table 2 summarizes Endress and Bonatti's experimental results. Experiments 6 and 7 were designed to control whether subjects considered either the initial or the final syllable to induce the rule, instead of both. Experiments 10 and 11 were designed to discount the possibility of phonological confounds in their Experiments 1 and 2, and Experiment 9 was designed to test whether a single mechanism can exploit TPs over both syllables and gaps. As our objective is to demonstrate that a single statistical mechanism can model the data that are relevant to the MOM hypothesis (rather than to develop a psychologically realistic model of speech segmentation), we ignore experiments 6–7 and 10–11. Experiment 9 is a variation of Experiment 3 where pure tones are used to surround test items. Running simulations where pure tones surrounding test items were represented, for example, by activation of an untrained input unit would add nothing substantial to the simulations herewith reported. Therefore, we address Experiment 9 in the general discussion.

Thus, we focus our discussion and simulations in the remainder of this paper upon the other experiments (1–5, 8, and 12–13), together with Experiments 1–5 by Peña et al. (2002a). The critical pattern in these experiments is that Peña et al. and Endress and Bonatti find a negative correlation between performance on abstract items and familiarization duration. Tables 3 and 4 list these experimental results from Peña et al. and Endress and Bonatti (which were ordered by their original experiment numbers in Tables 1 and 2) in order from

Table 2
Summary of Endress and Bonatti's experimental results (adapted from Endress and Bonatti, 2007)

Experiment	Stream	Familiarization Duration	Test Choice
1	Segmented	10 min	Class words over part words
2	Continuous	10 min	No preference between class words and part words
3	Segmented	2 min	Class words over part words
4	Segmented	30 min	No preference between class words and part words
5	Segmented	60 min	Part words over class words
6	Segmented	2 min	No preference between A_iC_jX and XA_iC_j
7	Segmented	10 min	Class words over $A_iX'A_j$ $C_iX'C_j$
8	Segmented	2 min	Words over rule words
9	Segmented	2 min	Class words over part words (surrounded by pure tones)
10	Segmented	2 min	Class words over part words
11	Continuous	2 min	No preference between class words and part words
12	Segmented	2 min	Rule words over class words
13	Continuous	10 min	Rule words over class words

Table 3
Summary of Peña et al.'s and some of Endress and Bonatti's experimental results with a *segmented* stream, in increasing order of familiarization duration

Experiment	Study	Familiarization Duration	Test Choice
5	Peña et al.	2 min	Rule words over part words
8	Endress and Bonatti	2 min	Words over rule words
3	Endress and Bonatti	2 min	Class words over part words
12	Endress and Bonatti	2 min	Rule words over class words
3	Peña et al.	10 min	Rule words over part words
1	Endress and Bonatti	10 min	Class words over part words
4	Endress and Bonatti	30 min	No preference between class words and part words
5	Endress and Bonatti	60 min	Part words over class words

Table 4
Summary of Peña et al.'s and some of Endress and Bonatti's experimental results with a *continuous* stream, in increasing order of familiarization duration

Experiment	Study	Familiarization Duration	Test Choice
1	Peña et al.	10 min	Words over part words
2	Peña et al.	10 min	No preference between rule words and part words
2	Endress and Bonatti	10 min	No preference between class words and part words
13	Endress and Bonatti	10 min	Rule words over class words
4	Peña et al.	30 min	Part words over rule words

shorter to longer familiarization times, with the goal of highlighting the relationship between the experimental results and the MOM hypothesis.

The results obtained by Peña et al. and Endress and Bonatti on *segmented* 2-min familiarization streams (Experiments 5, 8, 3, and 12 in Table 3) indicate preferences for words over rule words, for rule words over class words, and for class words over part words. Their results on segmented 10-min familiarization streams (Experiments 1 and 3 in Table 3) indicate preferences for class words and rule words over part words. Finally, Endress and Bonatti's results on segmented 30- and 60-min familiarization streams (Experiments 4 and 5 in Table 3) indicate no preference between class words and part words, and preference for part words over class words, respectively. We therefore observe both (a) a rank-order preference (words > rule words > class words > part words); and (b) a reversal in this order of preference between class words and part words (part words > class words) as familiarization exposures increase in time. On the other hand, the results by Peña et al. and Endress and Bonatti on *continuous* 10-min familiarization streams (Experiments 1, 2, 2, and 13 in Table 4) indicate preferences for words over part words, no preference between rule/class words and part words, and preference for rule words over class words. Finally, Peña et al.'s results on continuous 30-min familiarization streams (Experiment 4 in Table 4) indicate preference for part words over rule words. The overall reversal of the preferences observed (Tables 3 and 4) as we move from familiarization exposures of 2 min to those of 60 min reflects the negative correlation that underpins the MOM hypothesis.

Endress and Bonatti highlight the fact that participants' responses to class words exhibit a negative correlation between structural generalization and familiarization duration of much the same sort that Peña et al. (2002a) had found for rule words. Endress and Bonatti thus interpret the data (see Table 2) as showing that a dependency is initially induced between *classes of* items but degrades with further familiarization. This negative correlation is critical to the inference that there is MOM at work. Preference for class words over part words after 2 and 10 min of familiarization is taken as evidence that the participants have learned a class rule. Endress and Bonatti's reasoning is that familiarization with a segmented stream allows participants to focus upon the extraction of the underlying structure from the start, because they do not need to perform speech segmentation first. As familiarization duration increases, participants have more time to track the statistical relations that obtain between tokens in the input stream. Endress and Bonatti's interpretation is thus that an initially induced dependency between classes of items degrades with familiarization duration as it becomes overwhelmed by processing dependencies among speech elements.

1.2. A challenge for statistical learning

Endress and Bonatti (2007) consider whether a simple recurrent network (SRN; Elman, 1990) trained on a prediction task can account for the results of their experiments (Table 2). An SRN is a connectionist network that incorporates a context layer in addition to the input, hidden, and output layers found in a feed-forward network. The units in the context layer receive input from and direct output to the hidden layer. The weights from the hidden layer to the context layer are fixed so as to copy the contents of the hidden layer to the context

layer at each time step; the weights from the context layer to the hidden layer are trained by backpropagation, as are the weights from the input layer to the hidden layer and the weights from the hidden layer to the output layer. As in a feed-forward network, the units at each layer transform their inputs using an activation function. Over repeated presentations of input–output pairs, backpropagation reduces the amount of error at each output unit according to an objective function.

In particular, Endress and Bonatti are interested in whether a SRN can induce class words after being trained on segmented and continuous corpora of the sort employed in their Experiments 1–8. Endress and Bonatti purport to show that an associative connectionist network cannot account for this pattern. They report a set of studies with SRNs that they claim shows that a single mechanism like an SRN cannot account both for the preference for class words exhibited by humans in their experiments and for the negative correlation observed between class word induction and familiarization duration. In the remainder of this article, we report and discuss a set of SRN studies that *does* model the experimental results obtained by Peña et al. (2002a) and Endress and Bonatti (2007). Our goal is to model, using a single statistical mechanism, both the early preference hierarchy (words > rule words > class words > part words) and the reversal that obtains as familiarization durations are increased. As we shall see below, an interpretation that differs from Endress and Bonatti's is possible. To anticipate, participants may not be learning a class rule that, once acquired, gets overwhelmed with familiarization duration. Instead, as we shall argue, participants rule out non-acquired class words as familiarization continues. In what follows, we aim to show that a single statistical learning mechanism can in fact account for all the preference patterns in Tables 1 and 2, and for the negative correlation for both segmented and continuous corpora summarized in Tables 3 and 4.

2. Preliminary analysis of types of dependencies in Peña et al.'s corpus

As we saw earlier, Peña et al. (2002a) define a ''word'' in this series of experiments as a function of the TPs between syllables in A_i and C_i, respectively. This is for Peña et al.'s purposes of designing their experimental setting, but a number of different generalizations at several levels of abstraction may underlie the patterns of performance observed in the experiments (e.g., Seidenberg, MacDonald, & Saffran, 2002). Thus, in the same way as subjects may be tuning to the $A_i\_C_i$ rule in virtue of the TP of 1 between the first and the third syllables in the familiarization corpus, they may also be sensitive to generalizations such as $< A_iX$ *is always followed by* $C_i>$, or even $<A_i$ *is never followed by* $A_j>$, both generalizations with a TP of 1. Seidenberg et al. consider many other generalizations that subjects may rely on when processing the familiarization stream. Thus, in response to Experiments 1 and 2 by Peña et al., they claim that whereas words are supported, for example, by generalizations such as <*initial syllables begin with a stop consonant*>, <*final syllables begin with a stop consonant*>, <*continuant consonants occur word medially*>, among others, part words obtain their support from a smaller pool of generalizations. This might explain the observed preference for words in Experiment 1 of Peña et al. Similarly, the pool of generalizations

that is consistent with both rule words and part words is of the same size (see Seidenberg et al. for the details). This might help explain why subject preferences for either rule words or part words converge in Experiment 2 of Peña et al. It is of course not obvious that participants must be sensitive to all possible generalizations and to their corresponding TPs. Different types of information may be weighted differently, and the underlying empirical question in dispute is precisely what sources of information people track as they acquire a new language. In fact, a number of experiments by Endress and Bonatti (2007) were designed to control for some of these potential sources of information (see general discussion). In our view, however, and while acknowledging that these control experiments have served to discard alternative hypotheses such as the possibility of phonological confounds, there is no principled reason to exclude the possibility that subliminal segmentation gaps can be exploited statistically. The mere fact that these gaps are subliminal does not prevent them from carrying potentially relevant information. It simply means that their presence is not available to conscious access. As a matter of fact (as Peña et al. well observe), they must carry the critical piece of information for the mastery of structural induction, because the inclusion of the gaps is the only difference between Experiment 2 (where part words are preferred), and Experiment 3 (where rule words are favored) (Table 1).

Table 5 collects a number of generalizations, with their TPs, for a legal sequence of syllables in the familiarization corpus subject to the constraints that the TP between any A_i and the following C_i is 1.0, between any A_i and an intermediate X, and between an X and the final C_i, are each 0.33, and between any C_i and the next word's first syllable is 0.5. Table 5 reflects some of the statistical regularities in the corpora, adjacent as well as non-adjacent (Peña et al.'s Experiments 3 and 5, and Endress and Bonatti's Experiments 1, 3, 4, 5, 8, and 12), that participants have access to, and which may therefore explain their performance. Although Peña et al. consider a potential rejoinder according to which a single statistical mechanism may be responsible for the induction of the structural regularity in their Experiment 3 by tracking TPs over pauses as well as syllables, they dismiss that alternative

Table 5
Some potential generalizations about adjacent and non-adjacent syllables based on the familiarization sequence … $\#A_iXC_i\#A_jYC_j\#A_kZC_k\#$ … (pauses represented by ''#'')

No.	Generalization	TP
1	''#'' predicts ''A_i''	0.33
2	''A_i'' predicts ''X''	0.33
3	''X'' predicts ''C_i''	0.33
4	''C_i'' predicts ''#''	1
5	''#'' predicts ''X,'' after one intervening item	0.33
6	''A_i'' predicts ''C_i,'' after one intervening item	1
7	''X'' predicts ''#,'' after one intervening item	1
8	''C_i'' predicts ''A_j,'' after one intervening item	0.5
9	''#'' predicts ''C_i,'' after two intervening items	0.33
10	''A_i'' predicts ''#,'' after two intervening items	1
11	''X'' predicts ''A_j,'' after two intervening items	0.33

Note. TP, transitional probabilities.

in a footnote (note 27). The rejoinder is the idea that participants might have computed TPs to and from the pauses as well as to and from the audible syllables. In that case, they might estimate that a rule word at test (with the structure #$A_iX'C_i$#, where ''#'' stands for a subliminal segmentation gap) was more likely than a part word at test. (Part words can be of two types: ''type 12'' part words consist of items having the form C_iA_jX, whereas ''type 21'' part words consist of items having the form XC_iA_j.) The relevant TPs for a rule word and for part words of type 12 and type 21 (with the structure #CiAjX #and #XCiAj #, respectively) are shown in Fig. 1.

As Fig. 1 illustrates, whereas a rule word of the form #$A_iX'C_i$# can in principle be supported by five different generalizations (1, 4, 6, 9, and 10 in Table 5), part words of type 12 (#C_iA_jX#) and of type 21 (#XC_iA_j#) are exclusively supported by generalizations 2 and 3, respectively (all other adjacent and non-adjacent TPs among syllables are zero). Indeed, were we to consider just adjacent TPs, rule words would still be supported by two different generalizations (1 and 4 in Table 5), whereas TPs backing part words of either type reduce to 0.33 (generalizations 2 and 3 in Table 5). Thus, one might suppose that participants would still prefer rule words to part words based strictly on adjacent TPs.

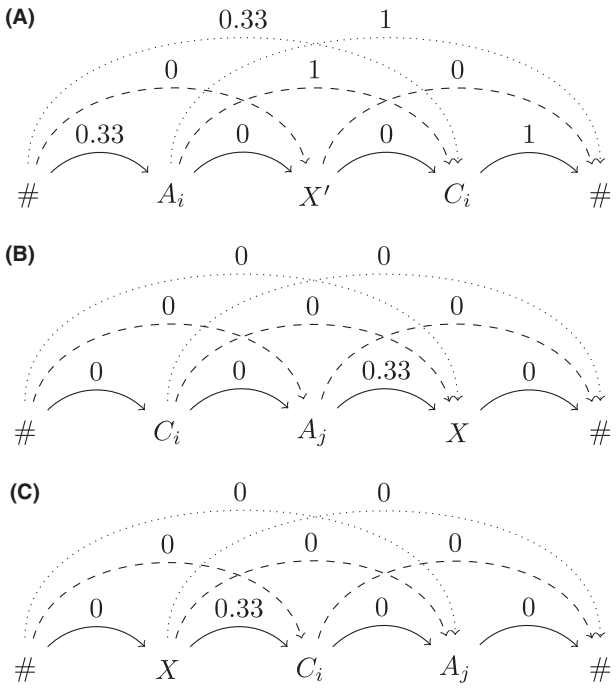

Fig. 1. Adjacent and non-adjacent transitional probabilities for (A) a rule word having the form #$A_iX'C_i$#; (B) a part word of type 12 having the form #C_iA_jX#; and (C) a part word of type 21 having the form #XC_iA_j#. Solid arrows indicate transitions between adjacent items. Dashed arrows indicate transitions between non-adjacent items separated by one. Dotted arrows indicate transitions between non-adjacent items separated by two. Each arrow is labeled with the corresponding transitional probability.

Peña et al. attempted to rule this hypothesis out by performing a control experiment in which participants were tested with test items that consisted of part words of type 21 *including* the internal pauses, that is, items having the form #XC_i#A_j#. The relevant TPs are shown in Fig. 2. In this case, a part word of type 21 with an internal gap (#XC_i#A_j#) would be supported by generalizations 1, 3, 4, 7, 8, and 11 in Table 5, or by generalizations 1, 3, and 4, were we to focus exclusively upon TPs between adjacent items. Thus, if participants were using only TPs between adjacent items in calculating their preferences between test items, we might expect them to prefer the part words of type 21 including the internal pauses (supported by generalizations 1, 3 and 4) over the rule words (supported by generalizations 1 and 4). Peña et al., on the contrary, report that participants still prefer rule words to part words even when the part words are presented with internal pauses.

However, their analysis ignores non-adjacent TPs. They consider the prediction that participants would choose rule words (#$A_iX'C_i$#) over part words (#C_iA_jX#), once we consider ''probabilities over syllables, pauses, and absence of pauses in the stream and the test items,'' since ''[t]ransitional probabilities *between adjacent elements* would favor rule words over part words'' (note 27, p. 607, emphasis added). No reason is offered as to why only adjacent TPs should be computed. Their conclusion is based upon the *assumption* that a statistical learning mechanism can only be sensitive to adjacent TPs. However, there is no reason to believe that such mechanisms cannot be sensitive to non-adjacent regularities (see, e.g., Gómez & Maye, 2005; Newport & Aslin, 2004). In fact, there is also a fairly clear literature demonstrating that recurrent networks can induce grammars from examples of context-free and context-sensitive languages; grammars that are precisely of a form in which there are long-distance dependencies (see, e.g., Boden & Wiles, 2000; Chalup & Blair, 2003). This is especially so given that Peña et al.'s experimental setting was precisely designed by constructing a lexicon mainly characterized in terms of *non-adjacent* TPs; probabilities which, as they themselves acknowledge, are the cornerstone of the segmentation task in their Experiment 1: ''[We] explore whether participants can segment a stream of speech by means of *nonadjacent* transition probabilities, and we also ask whether the same computations are used to promote the discovery of its underlying grammatical structure'' (pp. 604–605; emphasis added).

We may then ask: which test items would participants choose if they were computing TPs over *both* adjacent and non-adjacent items? As we have just seen, rule words may in principle be supported by five different generalizations (1, 4, 6, 9, and 10 in Table 5). Part words

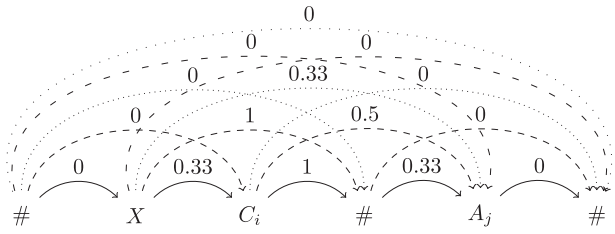

Fig. 2. Adjacent and non-adjacent transitional probabilities for a part word of type 21 with an internal gap, that is, an item having the form #XC_i#A_j#.

with an internal pause, on the other hand, are supported by generalizations 1, 3, 4, 7, 8, and 11. However, although part words with an internal pause are supported by more generalizations than rule words, the number of generalizations with a TP of 1.0 is bigger in the case of rule words (generalizations 4, 6, and 10 in Table 5, as opposed to generalizations 4 and 7 in the case of part words with an internal pause). Thus, if participants are in fact computing TPs over both adjacent and non-adjacent items, then there is no reason not to expect them to prefer rule words over part words with an internal pause, exactly as Peña et al. report that they do. It is not obvious that participants must be sensitive to all generalizations in virtue of their corresponding TPs. Thus, what the current discussion shows is that *the possibility* that statistics is behind it cannot be excluded *in principle* on the basis of an alleged statistical inferiority on the part of rule words as opposed to part words with an internal pause.

Our point is that statistical computations based on non-adjacent TPs of the sort that are exploited in speech segmentation may be used in order to induce existing grammatical regularities in the speech stream. In order to empirically demonstrate these claims, and with an eye to undermining one argument for the MOM hypothesis, we ran a series of connectionist simulations that illustrate the exploitation of statistically driven information. The simulations were conducted in two separate studies. In the first study, we aim to determine whether SRNs can exhibit the patterns of preference in Table 3 when trained on a corpus that contains subliminal gaps (Peña et al.'s Experiments 5 and 3, and Endress and Bonatti's Experiments 8, 3, 12, 1, 4, and 5; see Table 3). In the second study, we aim to determine whether SRNs can exhibit the patterns of preference in Table 4 when trained on a corpus that does not contain subliminal gaps (Peña et al.'s Experiments 1, 2, and 4, and Endress and Bonatti's Experiments 2 and 13; see Table 4).

3. Simulation studies

In our SRN studies, the familiarization corpus consisted of the same strings of syllables used by Peña et al. (2002a) and Endress and Bonatti (2007). In particular, syllables were represented by pairwise orthonormal nine- or ten-dimensional binary vectors (depending on whether segmentation gaps were included in the familiarization). The familiarization corpora were as close as possible to those used by Endress and Bonatti while still respecting the constraints described by Peña et al. The specific word classes used by Peña et al. (2002a) were:

$i = 1$: pu . . . ki

$i = 2$: be . . . ga

$i = 3$: ta . . . du

and the filler syllables were:

li

ra

fo

Thus, the A_1XC_1 family consists of the words *puliki*, *puraki*, and *pufoki*; the A_2XC_2 family consists of the words *beliga*, *beraga*, and *befoga*; and the A_3XC_3 family consists of the words *talidu*, *taradu*, and *tafodu*. To create the 10-min familiarization stream, 100 tokens of each of the nine words in Peña et al.'s lexicon were randomly concatenated, subject to two constraints: (a) a word of a family could not be followed by another word of the same family; and (b) two words could not be adjacent if they had the same intermediate syllable.

In generating our familiarization corpus, we did not use the constraint that there must be exactly 100 words of each type. Rather, for each set of simulations, we pseudorandomly generated 900 words according to the other constraints (see Appendix A for details and explanation).

We created seven test corpora to investigate the predictions of Peña et al. and Endress and Bonatti: (a) words (A_iXC_i); (b) part words of type 12 (C_kA_iX); (c) part words of type 21 (XC_iA_j); (d) rule words ($A_iX'C_i$); (e) class words ($A_iX'C_j$); (f) part words of type 12 that include internal gaps of the sort considered in footnote 27 of Peña et al. ($C_k\#A_iX$); and (g) part words of type 21 that include internal gaps of the sort considered in footnote 27 of Peña et al. ($XC_i\#A_j$).

3.1. Study 1

3.1.1. Method

Like Endress and Bonatti, we used an SRN (Elman, 1990). The syllables were coded as 10-bit pairwise orthonormal binary vectors (a ''1-of-c'' encoding), with the 10th bit representing a gap. We used the softmax activation function at the output layer combined with the cross-entropy objective function (e.g., Bishop, 1995). Based on the results of preliminary studies (reported in Laakso & Calvo, 2008), we set momentum to 0 and used 54 hidden units. Presenting a word to the network consisted of sequentially presenting each of its three syllables, followed by a gap. Networks had the same number of output units as input units and were trained to predict the next syllable (or gap) from each syllable (or gap) presented as input (more on the role of gaps below).

Trained networks were tested on five item types: training words ($N = 9$), part words of type 12 ($N = 18$), part words of type 21 ($N = 18$), rule words ($N = 12$), and class words ($N = 12$). The part words used for testing included internal gaps.

In their simulations, Endress and Bonatti considered three possibilities for representing and training on the segmentation gaps: (a) representing the gaps with a vector of length 0, that is, as the vector <0, 0, 0, 0, 0, 0, 0, 0, 0, 0>, and training the SRN to predict the gap just as it would predict a syllable; (b) representing the gaps with a vector of length 0 and training the SRN to predict the syllable immediately after the gap; and (c) representing the gaps by an extra unit and training the SRN to predict the syllable immediately after the gap. In selecting their representational scheme, however, Endress and Bonatti rely upon some of Peña et al. (2002a)'s data (in particular, the data Peña et al. report in footnote 27 of their paper that they claim suggests participants' performance does not depend on the gap being present in the test items). Endress and Bonatti interpret

this to mean that the best way to model participants' representations of the test items is by using a network that always predicts the next syllable, ignoring the gaps. In our preliminary work (Laakso & Calvo, 2008), we considered representing silences by an extra symbol. Moreover, in addition to testing networks on items that either contained no gaps or contained gaps at the beginning of test items, we also tested networks on items that contained gaps before the A syllables (rule words and class words began with a gap, part words of type 12 contained a gap between the first and the second syllables, and part words of type 21 contained a gap between the second and the third syllables). As we reported, even networks tested with gaps within part words exhibited a preference for rule words over part words, modeling the human behavior in the control experiment reported in footnote 27 of Peña et al. (2002a).

For the purposes of direct comparison with Endress and Bonatti's results, however, in the experiments reported here we tested the networks, as they did, by recording the network output for the second syllable of the test items (i.e., the network's prediction of what the third item would be) and then comparing the network output with the actual third syllable of the test item using the cosine similarity measure. (The cosine similarity measure has a value of 1 when two vectors point in the same direction, a value of −1 when they point in opposite directions, and a value of 0 when they are orthogonal.) That is, the cosine similarity measure was recorded between the third syllable of the test item and the network output activation in response to the second syllable of the test item. We performed this procedure for all of the test items, thereby recording network responses to all legal continuations of the first two syllables for each test item type.

Fifty networks with different random starting weights were trained in order to simulate individual differences. After every 10 epochs of training (each epoch consisted of a single presentation of all 900 words in the familiarization corpus), the performance of each network was measured and recorded. Training was stopped after 300 epochs.

3.1.2. Results

The results are shown in Fig. 3, which depicts cosine similarity values for words, part words, rule words, and class words over the course of 300 epochs of training averaged across our 50 network ''subjects.'' For convenience of exposition, we focus on performance after 50, 70, 100, and 200 epochs of training, as indicated by the vertical lines in the figure. For the same reason, we use the shorthand that the networks ''prefer'' test items of one type to test items of another type to stand for the cumbersome expression that the mean cosine similarity between the network outputs and the targets for the first test item type is greater than the mean cosine similarity between the network outputs and the targets for the second test item type.

After 50 and 70 epochs of training, the networks prefer words to rule words, rule words to class words, and class words to part words. After 100 epochs of training, the networks show no preference between class words and part words. After 200 epochs of training, the networks prefer part words to class words.

Fig. 3 shows only a single line for part words (PW). Fig. 4 shows the same results except that the data for part words of type 12 and part words of type 21 are depicted separately.

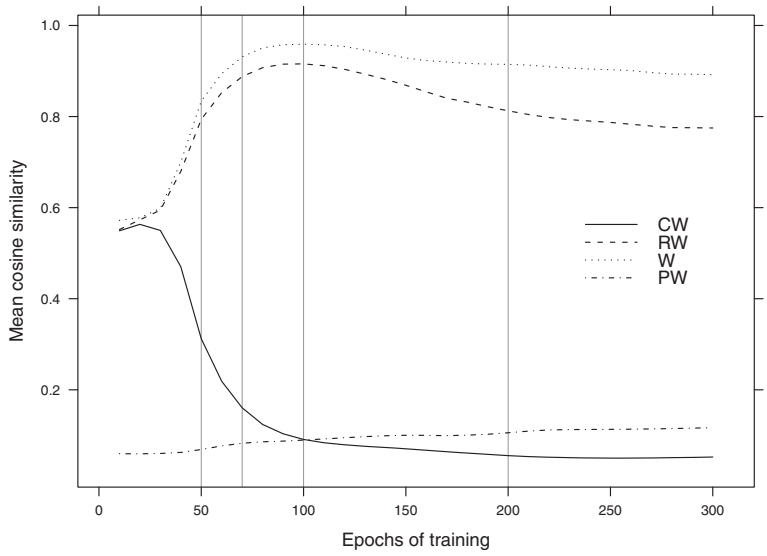

Fig. 3. Mean cosine similarity values for 50 networks trained and tested with gaps. CW, class words; PW, part words; RW, rule words; W, words.

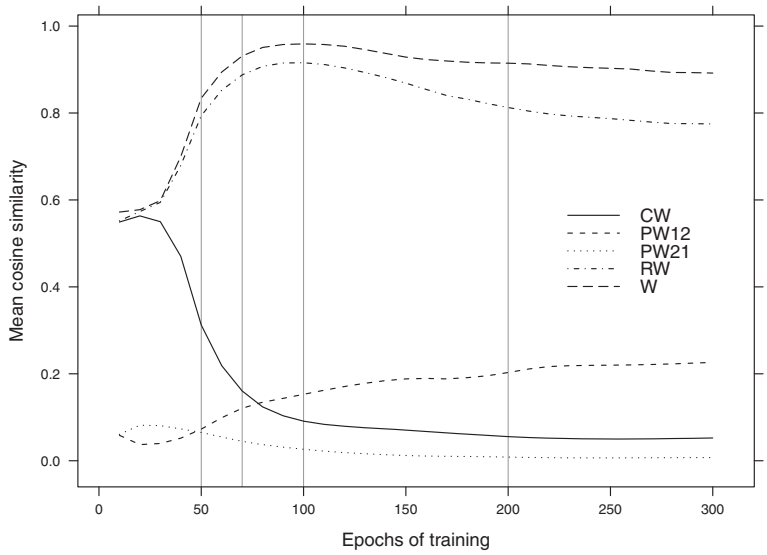

Fig. 4. Mean cosine similarity values for 50 networks trained and tested with gaps, with part word types shown separately. CW, class words; PW12, part words of type 12; PW21, part words of type 21; RW, rule words; W, words.

Because the plotted values are the same for all test item types except for part words, we present results here only for part words of each type in comparison to each other and to class words.

After 50 epochs of training, the networks prefer class words to part words of either type, showing no preference for part words of one type over the other. After 70 epochs of training, the networks prefer class words to part words of type 12, which in turn are preferred to part words of type 21. After 100 and 200 epochs of training, the networks prefer part words of type 12 to class words and class words to part words of type 21. Detailed descriptive and inferential statistics for Study 1 are presented in Appendix B.

3.1.3. Discussion

The results of our Study 1 (Fig. 3) accurately model the behavior of human participants in Peña et al.'s Experiments 5 and 3, and Endress and Bonatti's Experiments 8, 3, 12, 1, 4, and 5; see Table 3. Specifically, the results by Peña et al. and Endress and Bonatti on *segmented* 2-min familiarization streams (Experiments 5, 8, 3, and 12 in Table 3) indicate preferences for words over rule words, for rule words over class words, and for class words over part words (compare network performance in Fig. 3 after 50 epochs of training). Their results on segmented 10-min familiarization streams (Experiments 3 and 1 in Table 3) indicate preferences for class words and rule words over part words (compare network performance in Fig. 3 after 70 epochs of training). Their results on segmented 30-min familiarization streams (Experiment 4 in Table 3) indicate no preference between class words and part words (compare network performance in Fig. 3 after 100 epochs of training). Finally, their results on segmented 60-min familiarization streams (Experiment 5 in Table 3) indicate preference for part words over class words (compare network performance in Fig. 3 after 200 epochs of training).

It is important to point out that the divergence between the part word types in Fig. 4 is different from some behavioral results that Endress and Bonatti briefly report. As Fig. 4 shows, part words of type 12 and part words of type 21 fare differently with respect to class words in our simulations. Part words of type 12 (as shown in Fig. 4) show the same qualitative pattern as the average of all part words (as shown in Fig. 3). That is, they are dispreferred to class words after fewer epochs of training but preferred to class words after more epochs of training. However, part words of type 21 (as shown in Fig. 4) do not show the same qualitative pattern as the average of all part words (as shown in Fig. 3). In particular, our networks never prefer part words of type 21 to class words. Having observed a similar divergence in one of their network simulations, Endress and Bonatti wrote that ''the network predicts a difference between how part-words of type 12 and type 21 will stand the comparison with class-words; yet, in none of our experiments have we observed it'' (pp. 282–3).

Does this difference between the way that the networks perform and the behavioral data undermine our argument against the MOM hypothesis? We think not, for three reasons: (a) the human behavioral data are incomplete; (b) there could be several reasons for the difference between the network simulations and the human data, none of which are relevant to our main thesis; and (c) the very question whether the behavioral data and the simulated data match in this case is irrelevant to our thesis to begin with. In the following paragraphs, we address these reasons in turn.

First, the human behavioral data are incomplete. Endress and Bonatti write that they have not observed such a difference in their experiments, but they do not report the relevant data in detail. For one thing, although they report the standard deviation for participants' preference for class words, they do not report standard deviation for participants' preference for part words, whether combined or separate. (We explain in the next paragraph why this is important.) For another, none of the comparisons that Endress and Bonatti (2007) report is a direct comparison between part words of type 12 and part words of type 21. Instead, they report comparisons between part words of each type and class words. Thus, with regard to their Experiments 1–3 and 9–10, Endress and Bonatti report that there ''was no difference between the part-word types against which the class-words were tested'' (pp. 255, 256, 258, 269, 273). The assumption that part words of different types are indistinguishable is an inference from the experimental comparison between part words and class words to a prediction about the status of part word *types*; an inference whose validity we question. A sufficiently powerful test of the hypothesis that participants will respond differently to part words of different types is therefore needed. The appropriate experiment to test the hypothesis would compare part words of type 12 and part words of type 21 directly by forcing participants to choose between them at test. Our model suggests that participants might reliably prefer one of the part word types to the other, and that the preference might reverse after extended familiarization.

The second reason that the difference between the way that the networks perform on part words versus class words and the behavioral data does undermine our argument against the MOM hypothesis is that there could be several reasons for the difference between the network simulations and the human data, none of which would impugn our argument against MOM. One possibility is that the behavioral experiments simply did not have the statistical power to detect the small differences that actually do exist. Because Endress and Bonatti do not report the variance in preferences for the respective part word types, we do not know whether their failure to find a difference in part word types is merely due to lacking the statistical power that would be necessary to detect such differences. The amount of variance in the human data generally is quite high (presumably due to a plethora of irrelevant performance factors) compared to the amount of variance in the network data (because the networks are much simpler mechanisms than human beings), and the number of participants that Endress and Bonatti used in their human experiments (approximately 20 for each experiment) is fewer than the number of network participants that we trained (50).

Another possible reason that the network simulations may appear different from the human data in this respect is that the human preference structure may be non-metric. Human similarity judgments in many domains are non-metric (e.g., Tversky, 1977), but standard SRNs—whose representational mechanisms are fundamentally Euclidean—cannot exhibit non-metric preference structures. Admittedly, this points to one way in which the SRN simulations may fail to accurately model all of the human data. However, we must remember that the goal here is to model just the set of behavioral experiments that have been presented as evidence for the MOM hypothesis, in order to demonstrate that in principle a single statistical mechanism can account for the data, not to present a psychologically realistic model of human artificial language learning.

Yet another possible reason for the divergence observed between the two types of part words is the particular representational scheme that Endress and Bonatti chose to model participants' representations of the test items, which we copied for the purposes of direct comparison with Endress and Bonatti's results. Fig. 1b,c above, where adjacent and nonadjacent TPs for a part word of type 12 and a part word of type 21 were shown, may hint at the reason for the divergence between the patterns of part words of different types in Fig. 4. In the case of part words of type 12, the TP between the second and the third syllable is 0.33 (Fig. 1b). By contrast, the TP between the second and the third syllable in the case of part words of type 21 is 0 (Fig. 1c). Fig. 4 records the cosine similarity measure between the third syllable of the test item and the network output activation in response to the second syllable of the test item.

Having said all this, the third reason that the difference between the way that the networks perform on part words versus class words and the behavioral data does not undermine our argument against the MOM hypothesis is that the difference is irrelevant to the question at hand. As we noted in the introduction, our goal is to model the negative correlation between performance on abstract items and familiarization duration that Endress et al. have taken as evidence for the MOM hypothesis (i.e., the primary data listed in Tables 3 and 4). We have not claimed to offer a psychologically realistic model of every aspect of the human data. Therefore, the fact that there are minor discrepancies between the network performance and the human behavioral performance on an incidental measurement (part words of different types vs. class words) is tangential. It does not undermine our point that the networks model the most important effects in the human behavioral data, the ones that have been taken as the primary evidence for the MOM hypothesis.

Nevertheless, in addition to modeling the behavior of human participants in Peña et al.'s and Endress and Bonatti's experiments, our simulations reveal something important about the preference for class words. As Figs. 3 and 4 show, class rules do not appear to have been learned in the first place.

Endress and Bonatti interpret human participants' preference for class words over part words after 2 min and 10 min of familiarization as evidence that the participants have learned a class rule. Endress and Bonatti conjecture that when confronted with a segmented familiarization stream, participants may be freed from the burden of having to parse the input to extract first the constituents, as in the case of a continuous stream. In this way, they are able to focus from the start upon the underlying structure itself. With longer familiarization, generalization is overwhelmed as participants have more time to track the statistical relations that obtain between items in the input stream. The interpretation of Endress and Bonatti is thus that an initially induced dependency that takes place between *classes of* items degrades with familiarization duration. But Fig. 3 makes it clear that another interpretation of these results is possible: Perhaps the discovery of structural regularities is only *apparent* and people are not learning a class rule at all but only ruling it out rather slowly. That is what our networks seem to be doing—although the networks disprefer part words right from the beginning, it takes them some time (approximately 100 epochs—see Fig. 3) to learn that class words are no better.

Now why would participants be slower to develop a dispreference for class words than to show a dispreference for part words? Fig. 5 shows adjacent and non-adjacent TPs for a class word having the form #$A_iX'C_j$#. Note that, whereas class words can in principle be supported by two different generalizations among adjacent items (generalizations 1 and 4 in Table 5), and by two other generalizations among non-adjacent items (generalizations 9 and 10 in Table 5), part words of type 12 (#C_iA_jX#), and of type 21 (#XC_iA_j#), are only supported by one generalization among adjacent items (generalizations 2 and 3 in Table 5, respectively), and by none among non-adjacent items. An associative explanation may thus underlie the diverging time spans of dispreference between class words and part words shown in Fig. 3. We suggest that much the same might underlie human participants' performance in Endress and Bonatti's experiments. That is, people may be ruling out an unlearned class rule rather slowly. In fact, taking into account that the non-adjacent TP between the first and the third syllable is lost in the case of class words, it is difficult to grasp how participants might possess such powerful generalization machinery that is triggered when released from having to parse the corpus, as Endress and Bonatti suggest. More recently, Endress and Mehler (2009) have questioned the induction of classes. Instead, they now favor an edge-based mechanism that tracks the positional information of syllables in beginning and end position. Insofar as such an edge-based mechanism is non-statistical, the evidence they report still supports the MOM hypothesis (we address this other possibility in the general discussion).

A final caveat is in order. One might question the sense in which epochs of training in artificial neural networks and familiarization duration with human participants relate to each other (recall that network performance was probed every 10 epochs, and connection weights frozen after 300 epochs). Indeed, it is not clear that there must be a linear relation between epochs of training and familiarization duration. Our key point is that the networks do reproduce the initial patterns of preference, and that such a preference does reverse in subsequent epoch intervals as a result of an increased learning of the prediction task. The MOM hypothesis capitalizes on an observed negative correlation between the extraction of structural regularities and familiarization duration. The longer the duration of the continuous familiarization stream, the stronger the preference for part words over rule words. On the contrary, a very short familiarization with a segmented stream allows for the induction of

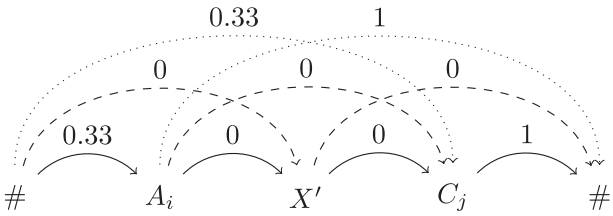

Fig. 5. Adjacent and non-adjacent transitional probabilities for a class word having the form #$A_iX'C_j$#. Solid arrows indicate transitions between adjacent items. Dashed arrows indicate transitions between non-adjacent items separated by one. Dotted arrows indicate transitions between non-adjacent items separated by two. Each arrow is labeled with the corresponding transitional probability.

the rule (preference of class and rule words). Thus, Endress and Bonatti predict that a preference for class words will decrease over longer familiarization durations. Demonstrating that, with sufficient training, networks can also show a reversal (coming to prefer part words over class words) is thus critical to the debate. Such a reversal is exactly what we find in Study 1. The MOM hypothesis ignores the possibility that subliminal segmentation gaps can be exploited statistically, as the present results with SRNs illustrate.

3.2. Study 2

The previous simulation used training data matching the familiarization stimuli used in Peña et al.'s Experiments 3 and 5, and Endress and Bonatti's Experiments 1, 3, 4, 5, 8, and 12 (Table 3). However, an important aspect of Peña et al.'s and Endress and Bonatti's results is the fact that their experiments show that participants cannot learn the so-called abstract rules without the segmentation gaps. How then do our SRNs fare when trained on a familiarization corpus that does not include the gaps (Table 4)? The results obtained by Peña et al. and Endress and Bonatti's on *continuous* 10-min familiarization streams (the first four experiments in Table 4) indicate preference for words over part words, no preference between rule/class words and part words, and preference for rule words over class words. Finally, Peña et al.'s results on continuous 30-min familiarization streams (Experiment 4 in Table 4) indicate preference for part words over rule words. Study 2 attempts to model the pattern of performance in Table 4.

3.2.1. Method

Study 2 was identical to Study 1, except that the networks were trained and tested without gaps. Thus, the syllables were coded as nine-bit pairwise orthonormal binary vectors. As previously, the networks used the softmax activation function at the output layer, the cross-entropy objective function, momentum of 0 and 54 hidden units.

Presenting a word to the network consisted of sequentially presenting each of its three syllables. Networks had the same number of output units as input units and were trained to predict the next syllable from each syllable presented as input. The testing procedures were the same as in Study 1, except that the test items did not contain gaps. As in Study 1, 50 networks were trained in order to simulate individual differences. After every 10 epochs of training, the performance of each network was measured and recorded. Training was stopped after 300 epochs.

3.2.2. Results

The results are shown in Fig. 6, which depicts cosine similarity values for words, part words, rule words, and class words over the course of 300 epochs of training averaged across our 50 network ''subjects.'' For convenience of exposition, we focus on performance after 50 and 200 epochs of training, as indicated by the vertical lines in the figure. For the same reason, we again use the shorthand that the networks ''prefer'' test items of one type to test items of another type to stand for the cumbersome expression that the mean cosine similarity between the network outputs and the targets for the first test item type is greater

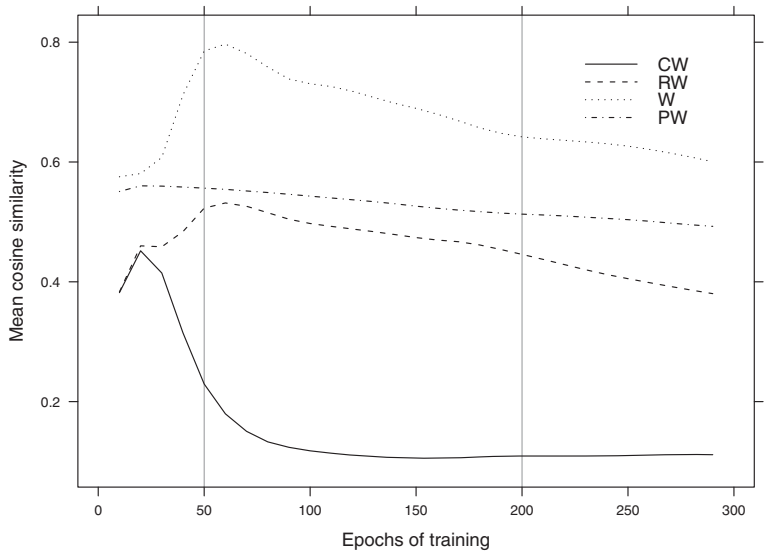

Fig. 6. Mean cosine similarity ratings for 50 networks trained and tested without gaps. CW, class words; PW, part words; RW, rule words; W, words.

than the mean cosine similarity between the network outputs and the targets for the second test item type.

After 50 epochs of training, the networks prefer words to part words. In addition, they are indifferent to the choice between part words and rule words. Finally, they prefer rule words to class words. After 200 epochs, the networks prefer part words to rule words.

Fig. 7 shows the same results with part words of type 12 and part words of type 21 drawn separately. After 50 and 200 epochs of training, the networks exhibit a slight preference for part words of type 12 to part words of type 21 and for part words of type 21 to rule words. Detailed descriptive and inferential statistics for Study 2 are presented in Appendix C.

3.2.3. Discussion

The results shown in Fig. 6 match the experimental results in Table 4 with one exception: Although Endress and Bonatti found that human participants had no preference between class words and part words after 10 min of familiarization with a continuous stream (their Experiment 2—see Table 4 above), our networks prefer part words over class words throughout the course of training on inputs without gaps (Fig. 6). We must remember, however, the rationale of Endress and Bonatti's Experiment 2 in the context of testing the MOM hypothesis. Their concern was with whether participants could learn a class rule. They took the results of their Experiment 1 (showing a preference for class words over part words after 10 min of familiarization on a segmented stream—see Table 3 above) to show that participants could learn a class rule after familiarization with a stream containing segmentation cues. By contrast, they took the results of their Experiment 2 (showing no preference

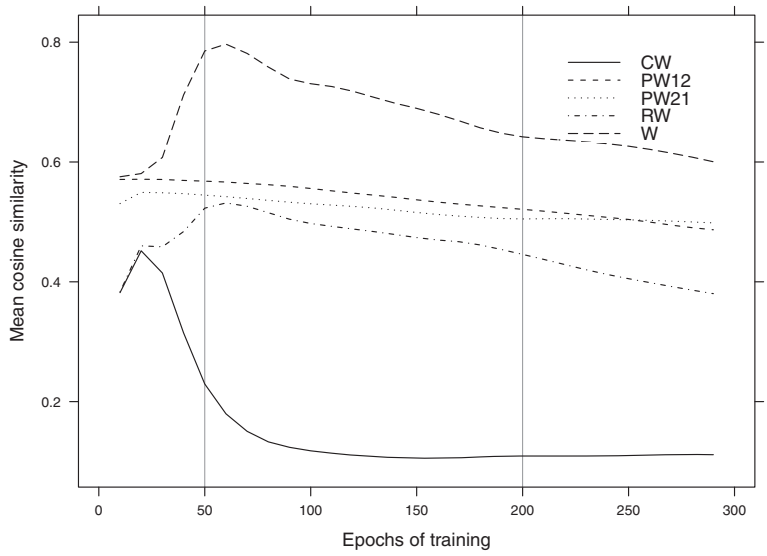

Fig. 7. Mean cosine similarity ratings for 50 networks trained and tested without gaps, with part word types shown separately. CW, class words; PW12, part words of type 12; PW21, part words of type 21; RW, rule words; W, words

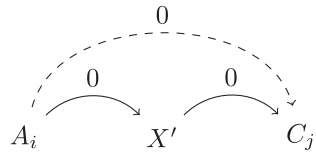

Fig. 8. Adjacent and non-adjacent transitional probabilities for a class word without surrounding gaps, that is, having the form $A_iX'C_j$. Solid arrows indicate transitions between adjacent items. The dashed arrow indicates the transition between non-adjacent items separated by one. Each arrow is labeled with the corresponding transitional probability.

between class words and part words after 10 min of familiarization on a continuous stream—see Table 4 above) to show that participants could not learn a class rule after familiarization with a stream not containing segmentation cues. Based on our network modeling results, however, we 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. (See Fig. 3 and the discussion in the results section of Study 1 above.) The same point applies to streams not containing segmentation cues as to streams that do contain segmentation cues: Networks rapidly develop a dispreference for class words (compare the lines for class words in Figs. 3 and 6), and we suggest that human participants may be doing the same. Fig. 8 shows adjacent and non-adjacent TPs for a class word without surrounding gaps, that is, a word having the form $A_iX'C_j$. Note that all of the TPs are 0. It is not clear why we would expect any learner to find such items familiar.

We therefore take the following lessons from the stimulations: (a) segmented inputs lead statistical learners to find both words and rule words familiar while finding class words and part words unfamiliar; and (b) unsegmented inputs lead statistical learners to find words familiar while finding rule words, part words, and class words unfamiliar. Other differences (i.e., the differences between part words, rule words, and class words in Fig. 6) may simply be uninteresting for participants as they are all ''below threshold'' for familiarity. Thus, the negative correlation between duration of familiarization and generalization of a class rule that is exhibited in the human experiments may be a red herring—rather than providing strong evidence for the existence of two different learning mechanisms, it may simply reflect overlearning of certain statistical patterns that are in any case irrelevant to the projection of a rule. Note how low the cosine similarity values are for part words and for class words after about 50 epochs of training in Fig. 6.

In summary, the experiments in Table 4 complement those of Table 3 insofar as no generalization obtains when the familiarization stream contains no gaps. The modeling results in Study 1 would be less significant were we to fail to model statistically the fact that no induction of rules obtains without the segmentation gaps. Studies 1 and 2 accurately, we conclude, model the experimental results that are relevant to the MOM hypothesis.

4. General discussion

Non-adjacent dependencies are an important test of any purported account of language acquisition based on statistical learning because they are pervasive in language, but it is not obvious how a statistical learning mechanism could acquire them. The negative correlation between performance on abstract items and familiarization duration (Tables 3 and 4) has been taken to underpin the MOM hypothesis. Initial preference for class words over part words is taken as evidence that the participants have learned a class rule. Endress and Bonatti's reasoning was that familiarization with a segmented stream allows participants to focus upon the extraction of the underlying structure from the start. However, with further familiarization, participants can track the existing statistical relations between stream chunks. Thus, according to Endress and Bonatti, an initially induced dependency between classes of items degrades with familiarization duration as it becomes overwhelmed by processing dependencies among speech elements. In this way, according to the MOM hypothesis, although an associative mechanism suffices to segment speech, structural generalization requires a rule-following mechanism. Both an associative and a non-associative mechanism are thus needed to analyze speech.

In this section we (a) compare our analysis vis-à-vis others; (b) discuss the source of the discrepancies between the work of Endress and Bonatti and ours; (c) address issues of implementation; (d) assess further evidential basis on behalf of MOM hypotheses; and close up by (e) considering the distinction between types and tokens of mechanisms.

4.1. Our analysis vis-à-vis others

We are not the first to have criticized the MOM framework. Previous criticisms that call into question the dual-mechanism account that underpins the MOM framework of Endress and Bonatti (2007) have been raised by Newport and Aslin (2004), Onnis, Monaghan, Richmond, and Chater (2005), Perruchet, Tyler, Galland, and Peereman (2004) and Perruchet, Peereman, and Tyler (2006), among others. So how does our analysis differ from these other criticisms?

Newport and Aslin (2004) and Onnis et al. (2005) explored the possibility of learning non-adjacent dependencies experimentally and via modeling, respectively. In particular, their work exploited the possibility of phonological/phonotactic confounds by focusing on previous linguistic knowledge of the subjects. The patterns of preference observed in Endress and Bonatti's experiments might be due to a match between statistical features of the artificial stream and the statistical distribution of words of their native language. Although it is possible that previously acquired knowledge helps explain the behavior of subjects in non-ecological experimental settings (it has been modeled for instance in the ''rule learning by infants'' debate by Seidenberg and Elman [1999a] in response to Marcus et al. [1999]), our simulations target different aspects of Peña et al. and Endress and Bonatti's results. We exploit the structure of the corpora themselves, a structure that contains, as we have argued, sufficient information, once subliminal segmentation gaps are included. The inclusion of silences being crucial in Peña et al. (2002a) and Endress and Bonatti (2007), we reasoned that it may be possible to track TPs not only among syllables but also among syllables and silences, regardless of the possibility of phonological/phonotactic confounds, and regardless of the possibility that previous cognitive pre-shaping (or, for that matter, any other source of information external to the experimental settings themselves) underpin the behavioral responses reported by Peña et al. and Endress and Bonatti.

Other replies, such as Perruchet et al. (2004, 2006), pinpointed ''deep methodological inadequacies.'' Courtesy of PARSER (Perruchet & Vinter, 1998), a model that generates streams exclusively based upon adjacent elements, Perruchet and colleagues targeted the possibility that participants may be tuning to parts of the stimuli (A_i or C_i) instead of generalizing the $A_i\_C_i$ rule itself (although see Bonatti et al., 2006, for a rejoinder). We need not repeat their exchanges here, but suffice it to say for present purposes that more recent work by Perruchet and Tillmann (2010) appears to confirm the view that the general-purpose learning principles underpinning PARSER may account for structural rule learning in artificial grammars. Perruchet and Tillmann (2010) argue that connectionist networks may also be able to account for the data they report, but at the expense of (ad hoc) sophistication. This claim goes beyond the present proposal, but despite the differences, the objective of this work is not to compare PARSER with connectionist networks, but rather to explore the possibilities of associative mechanisms, ultimately, a shared agenda.

With all that being said, Endress and Bonatti designed control experiments to deal with these and other concerns, one by one. Thus, for example, their Experiments 6 and 7 were designed to control whether participants considered either the initial or the final syllable to induce the rule, instead of both, and their Experiments 10 and 11 to discount the possibility

of phonological confounds in their Experiments 1 and 2. On the other hand, their Experiment 9, where class words and part words are both surrounded by pure tones during test, was designed to test whether a single mechanism could exploit TPs over both syllables and gaps. We are puzzled about what pure tones, rather than silences, are supposed to represent in Experiment 9. Endress and Bonatti (2007) claim: ''Because this manipulation obliterates TPs to and from silences surrounding test items, silences cannot play any crucial role to account for participants' preference for class-words found in Experiments 1 and 3. Experiment 9 demonstrates that associationist computations over syllables and silences alike cannot account for the preference for class-words'' (p. 270). We could have run simulations where pure tones surrounding test items were represented by activation of an untrained input unit, but we fail to see what the point would have been. After all, whether the bracket between test words be 25 ms silences (Endress and Bonatti, 2007), 1 s silences (Endress & Mehler, 2009; see below), pure tones, or whatever else someone might come up with, it does not change the sort of regularities involved within the test items. No such manipulation demonstrates that a mechanism that learns those regularities cannot be associative. Human participants may well automatically ignore the surrounding pure tones during test as nonlinguistic material, much as they must ignore any ''surrounding'' silences during test, since they are indistinguishable from silence before and after presentation of the test word. That is, when a human being performs a two alternative forced choice task with auditory stimuli, any stimuli that are ''preceded by silence'' are indistinguishable from those that are not (because, in such a task, all stimuli are preceded and followed by some silence anyway).

Once having controlled for, and discarded one by one, the alternatives proposed against the MOM hypothesis, Endress and Bonatti turned to connectionist simulations with SRNs. The idea was to demonstrate that there was no possibility for associative language learning, assuming that SRNs were, computationally speaking, representative of the associative mechanisms that may underlie language acquisition in humans.

4.2. The source of the discrepancies

In this article, we have taken up the challenge put forward by Endress and Bonatti and reported the results of a series of simulations, based on a single connectionist model, that accounts for the negative correlation observed in their earlier work. We attribute the divergence between our modeling work and theirs to three factors: (i) our familiarization corpora were generated in a superior fashion; (ii) we used different activation and objective functions; and (iii) our networks have more hidden units. We explain (i)–(iii) in turn.

First, Endress and Bonatti generated familiarization streams that ''contained 100 repetitions of each word, yielding 900 words in total'' (p. 279). However, as we explain in Appendix A, this is likely to have resulted in a biased familiarization corpus. To conform to Peña et al.'s original design, the familiarization stream must obey the following constraints: (a) a word of a given family cannot be followed by another word of the same family; and (b) two words that have the same intermediate syllable cannot be in adjacent position. So, for instance, *puliki* and *puraki*, and *puliki* and *beliga*, cannot be followed by each other on pain of violating constraints (a) and (b), respectively. However, if the corpus is built by first

taking 100 tokens of each word and then pseudorandomly concatenating them according to the constraints (a) and (b), it is very likely that a significant number of words near the end will be repetitions. (See Appendix A for the details and for additional causes of concern.) Thus, unlike Endress and Bonatti, we generated a familiarization corpus of 900 words selected pseudorandomly subject to the constaints (a) and (b), without being subject to the constraint that there should be exactly 100 tokens of each word. This procedure ensures that the familiarization stream does not end with a highly repetitive series of items.

The second difference between our simulations and those of Endress and Bonatti is in the activation and objective functions used in the neural network models. Endress and Bonatti do not report the activation functions or objective function used in their simulations. Given, nonetheless, their employment of standard, Elman-type, SRNs, we inferred that they used the logistic activation function, and an anonymous reviewer confirmed this guess. However, our simulations do not fall neatly within the standard sort of modeling that Endress and Bonatti appear to have in mind. (Endress and Bonatti consider work on SRNs by Altmann (2002), Christiansen and Curtin (1999), and Rodriguez (2001), versions that do not substantially depart from Elman's original model). There is a well-known issue with Endress and Bonatti's common choice (using sigmoid output units and the sum-squared error function) to train networks on problems where the target patterns are mostly zeros, as they are here. Such networks easily find a local minimum of the sum-squared error function by adjusting weights so that all output unit activations are close to zero. Moreover, because the delta term used in backpropagating sum-squared error involves a multiplication by the derivative of the activation function (the ''sigma prime term''), training slows down dramatically whenever the output approaches 0 or 1, regardless of the target value (because the derivative of the sigmoid approaches 0 in both cases). The preferred procedure for problems using a 1-of-c encoding is to use the softmax activation function at the output units combined with the cross-entropy objective function (e.g., Bishop, 1995). The softmax activation function causes the activations of the output units to always sum to unity, which is correct in the case of a 1-of-c encoding (a side effect is that one may treat output activations as the network's subjective assessments of the probability that each output unit codes for the right category on a given input pattern). In addition, using the cross-entropy objective function causes the sigma prime term to drop out of the calculation of delta values, ensuring that weight updates approach zero only as the activation value approaches the target value. Thus, in our simulations, we used the softmax activation function and the cross-entropy objective function.

The third difference between our simulations and those of Endress and Bonatti is the number of hidden units in the networks. On the basis of pilot studies (Laakso & Calvo, 2008), we determined that learning the familiarization stream worked best when networks had 54 hidden units. Thus, whereas the networks used by Endress and Bonatti had either 5 or 27 hidden units, hidden space in our case consists of 54 dimensions. Although hidden dimensionality was not an issue in their opinion (they report similar results on networks with 5 and 27 hidden units), this may simply be because their networks did not have enough hidden units to perform the task. Our model has more representational resources in hidden unit space due to our doubling the maximum number of dimensions they used.

In sum, in the light of our results it is clear that an argument is needed to defend Endress and Bonatti's claim that their ''results transcend [their] particular model, and that other associationist mechanisms will behave like SRNs with respect to our experiments.'' (p. 279). In fact, although we believe that our ability to simulate the data is due to a combination of (i)–(iii) above, determining the precise source of the discrepancies is secondary. The reason is that Endress and Bonatti's argument for the MOM hypothesis rests upon the assertion that no merely statistical mechanism can account for the empirical data (''we conclude that a single-mechanism hypothesis, as implemented by a SRN *or any associative device that extracts co-occurrences among items in the stream*, is not adequate to explain our data,'' p. 285; emphasis added); the MOM hypothesis thus depends upon a universal nonexistence claim. Our response has been to present *one* case of an associative device in which a merely statistical mechanism does account for the known empirical data; that is, we have presented an existence proof.

4.3. Issues of implementation

It is fair to say that concluding that algebraic manipulations do not take place based on the simple fact that a neural network accounts for the data is a *non sequitur*. The hidden premise that delivers the goods is that neural networks do not implement abstract relationships between variables in an explicit manner. Marcus (2001), for instance, has pursued this line of reasoning in the case of the so-called great past tense debate (Pinker & Ullman, 2002; Ramscar, 2002) and the debate over rule learning in infants, and concluded that only connectionist models that explicitly implement abstract relationships between variables can account for both the past tense and infant results. We need to know then which architectural features are behind our network's successful modeling of the experimental results of Peña et al. and Endress and Bonatti. If the required features include, for instance, the instantiation of variables, then a connectionist model of speech generalization will serve to back up the MOM hypothesis (by implementing abstract variables). We may put this somewhat differently: To warrant their claim that the MOM hypothesis is the only explanation for the data, Endress and Bonatti need to show that *only* connectionist networks that implement explicit rules in the form of abstract relationships between variables can account for their results.

The question then is: Which models implement rules? A clear case of implementation occurs when nodes in the architecture are used as variables. This is not the case with our SRNs, where the ''1-of-c'' encoding is used to represent the actual syllables that form the ''training'' words in familiarization (*pu* or *ki*), and not to represent word positional slots (A_i or C_i). Someone may argue that a different form of implementation takes place when an SRN is trained on a categorization task, where the teaching pattern for gradient descent learning is provided externally. Marcus (2001), for example, argues that the SRNs of Seidenberg and Elman (1999a) fall in this implementational category. As in the case of the encoding of variables, the argument would run, a rule is implemented in the non-ecological calculation of delta values in the form of a trainer that marks the output categories explicitly. In our simulations, however, the cross-entropy objective function deployed in a

self-supervised prediction task calculates delta values in an ecological, non-implementational manner. That is, no output categories are fed externally to the network as targets, because the target is the next input to the SRN, an input that stands for actual syllables and not variable slots.

There may be other ways in which the modeling results of an implementational connectionist network would support the MOM hypotheses. But it is noteworthy that Endress and Bonatti's failed attempt to model their own experimental results was carried out with SRNs. We may thus assume that their effort was directed toward a form of connectionism that they considered non-implementational. Because the only architectural difference between their model and ours is the employment of the softmax activation function at the output units combined with the cross-entropy objective function in an otherwise standard Elman net, we take it that they would agree that we have built a connectionist model that accounts for the negative correlation observed *without implementing* abstract relationships between variables in doing so.

4.4. Further evidential basis

Moreover, it is important to distinguish between positive evidence that supports the claim that algebraic mechanisms are fast and negative ''evidence'' that would support the same claim somewhat more indirectly. As Seidenberg et al. (2002) put it in response to Peña et al. (2002a), ''The evidence for rule learning is mostly negative: cases where learning occurs but there is no obvious statistical explanation. A theory explaining how rule learners arrive at exactly the correct generalizations given the complexities of their experience would represent substantial progress'' (p. 554). Allowing for the statistical generalization of rule words and the slower development of a dispreference for class words, together with the reversal of behavior we have modeled, we would like to see an explicit presentation of the nuts-and-bolts of a complex mechanism that would make it a more attractive alternative than our very simple model.

Interestingly enough, although Endress and Bonatti (2007) try to make progress by characterizing the operations that may underlie the negative correlation they observed,[1] in a more recent paper, Endress and Mehler (2009) themselves question the very possibility of class-learning in artificial grammars (see, e.g., Gerken, Wilson, & Lewis, 2005; Monaghan, Chater, & Christiansen, 2005; Redington et al., 1998). They argue instead in favor of an edge-based mechanism that tracks when the A_i and C_i elements are in the beginning and end positions. Someone might contend that, considering the way in which Endress and Mehler (2009) reinterpret Peña et al. (2002a) and Endress and Bonatti (2007), the work reported here becomes somewhat obsolete. However, although Endress and Mehler now agree that there is no class-learning as such, the new experimental evidence they report still serves, in their view, to back up the MOM hypothesis. Because the partial modifications that Endress & Mehler incorporate still serve to advocate a MOM view, our working hypothesis does bear directly upon their re-interpretation.

Briefly, Endress and Mehler (2009) designed a set of experiments aimed at assessing whether learning the positional information of A_i and C_i elements might account for Endress

and Bonatti's so-called class-learning in terms of an edge-based, non-statistical, mechanism. Their experiments are based on the same method as that used by Peña et al., and Endress and Bonatti, except that participants listen to sequences of pentasyllabic, instead of trisyllabic, artificial words. The idea, inspired by the sequential memory literature (e.g., Henson, 1998), is to test whether participants can generalize irrespective of the position of the A_i and C_i syllables (edge or middle position) as opposed to being able to generalize only when the A_i and C_i syllables are in edge position. Thus, under two experimental conditions, words can have the form A_iXYZC_i (edge condition), or the form XA_iYC_iZ (middle condition), respectively. If participants generalize exclusively after familiarization with edge-condition words, that may count as evidence for an edge-based mechanism. By contrast, Endress and Mehler reason, were participants to generalize after familiarization with both edge- and middle-condition words, that would count as evidence against an edge-based mechanism and in favor of some form of full-fledged class-learning ability that develops irrespective of edge-saliency, as in the mastery of the noun and verb categories of natural languages, which can be acquired regardless of their position.

Congenial with the results of Peña et al. and Endress and Bonatti with trisyllabic words, the results that Endress and Mehler report with pentasyllabic words are that participants fail to generalize after familiarization with a continuous stream under both conditions (edge and middle). However, they also fail to generalize after familiarization with a subliminally segmented stream (25 ms), under both conditions (edge and middle). But, finally, participants are able to generalize after familiarization with a *non*-subliminally segmented stream (1 s), although, crucially, *only* when the A_i and C_i syllables occur in edge position (A_iXYZC_i). That is, participants still fail to generalize when the A_i and C_i syllables occur in middle position (XA_iYC_iZ). These results, taken together, drive Endress and Mehler to suggest that participants are able to generalize courtesy of a non-statistical mechanism that operates by encoding the position of syllables in beginning and end position.

In our view, the results of Endress and Mehler can easily be accommodated within the general framework herewith advocated. Bluntly, why do we need a specialized edge-based mechanism to encode the positions of A_i and C_i syllables *once* positional cues (edges) are available? As we have seen before, a simpler explanation is available: a statistical mechanism that tracks TPs not only among syllables but also among syllables and silences. That is, instead of positing the existence of an edge-based mechanism, and a statistical mechanism that tracks TPs among syllables, the results reported in this article with trisyllabic words suggest that a single statistical mechanism may well be capable of tracking the different co-occurrence statistics in the edge and middle conditions with pentasyllabic words as well.

Moreover, it is important to highlight that the controversial subliminal character of the pauses used by Peña et al. and Endress and Bonatti are no longer an issue in the case of Endress and Mehler (2009). Participants fail to generalize with 25 ms gaps and are only able to do so when 1 s long segmentation pauses are inserted. In our view, then, there is no reason to exclude the possibility that such strongly marked co-occurrences cannot be tracked statistically, when we have seen that participants can do so even on subliminally segmented trisyllabic words. Based on our network modeling results, we have suggested that

participants may not be learning a class rule after all, 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 show a dispreference for part words. These results are directly applicable to pentasyllabic words: Time to develop a dispreference for CWs in edge and middle conditions may vary because of the differing weight of the TPs between silences and A_i and C_i syllables. A fast associative mechanism may thus account not only for Endress and Bonatti's results but also for Endress and Mehler's experiments with pentasyllabic words. Nevertheless, whether generalization on the new corpora could be accounted for by a single statistical mechanism, as we have shown for trisyllabic words, or by more than one statistical mechanism, remains a question for future research.

4.5. Types and tokens of mechanisms

Finally, it remains an issue whether statistical mechanisms cannot be fast. Perruchet et al. (2004) make a similar point when they note that: ''The assertion that associative learning proceeds slowly does not stand up to empirical observations. For example, some associative forms of learning have been shown to develop over one trial or so'' (p. 582). As discussed above, the defender of the MOM hypothesis needs to show that *only* connectionist networks that explicitly implement abstract relationships between variables can account for the Endress and Bonatti results. Presumably, the argument would be that the quick extraction of generalizations can be achieved only by a mechanism that implements abstract relationships between variables. However, we need not even try to qualify the claim that statistical mechanisms are necessarily slow. As Bonatti et al. (2006) acknowledge: ''We never denied that a theory based on statistical learning might account for fast learning.'' But they continue, ''however, the thesis that all learning can be explained by statistical computations is empty unless our critics can propose *a single mechanism* that is capable of *simultaneously* explaining (a) segmentation of words after exposure to a long familiarization (but not to a short familiarization) with a continuous stream; (b) extraction of structural information after a short familiarization with a discontinuous stream; and (c) failure to extract the same information after familiarization with any continuous stream'' (p. 319; emphasis added).

Now, granting that statistical learning may underlie fast learning, we can see that the challenge of Bonatti et al. (2006) is unjustified. The challenge resides not in accounting for (a)–(c) statistically, but rather in accounting for (a)–(c) *simultaneously* by means of *a single mechanism*. But they seem to be conflating types of mechanisms with specific mechanisms as tokens. In connectionist jargon, the challenge is then to obtain a single weight matrix through error-driven training such that all the knowledge that gets induced is fully distributed and superposed in the matrix. But why could not separate or partially overlapping statistical mechanisms be responsible for (a), (b), and (c)? Why, in short, is a single weight matrix needed where all knowledge is fully superposed? This rendering of the situation amounts to raising the challenge in terms of a single (token) mechanism. Bonatti et al. (2006) do not complain after all about the speed of one kind of learning over another, but rather about the number of token mechanisms that can play a role. The MOM hypothesis entails ''more than one *type* of mechanism,'' but nothing in their argument tells against the

possibility that more than one associative mechanism is in place. In short, someone may even wish to stick to the label ''MOM'' and defend on empirical grounds the existence of ''more than one (statistical) mechanism,'' such that *none* of the statistical mechanisms involved implements abstract relationships between variables. With that being said, the modeling results reported here account for the data in a non-implementational manner while meeting the added challenge of accounting for (a)–(c) above in one single weight matrix.

Overall, our results and discussion show that the MOM hypothesis is a less parsimonious explanation than a single type of mechanism (associative learning) hypothesis. Our model, we contend, meets this challenge: It explains the phenomena that Peña et al. and Endress and Bonatti report, and it is more parsimonious than their alternative. The single-mechanism hypothesis, moreover, generates further predictions for testing. As noted in Study 1, Endress and Bonatti have not reported experiments explicitly comparing preferences for part words of type 12 versus part words of type 21. Our prediction, based on associative learning principles and our simulations, is that such experiments will show a preference for part words of type 12 over part words of type 21 after 30 min or more of familiarization. Parsimony and novel testable predictions are an added value that in our view may end up tipping the balance against the dual-mechanism (MOM) hypothesis.

5. Conclusion

How many mechanisms are needed to analyze speech? According to the MOM hypothesis (Endress and Bonatti, 2007; Peña et al., 2002a), language learning is achieved by means of two mechanisms: a statistical mechanism that permits the learner to extract words from the speech stream, together with a non-statistical mechanism that is necessary for extracting higher level structure. We have presented a pair of neural network studies that show how statistics alone can support the discovery of structural regularities, beyond the segmentation of speech. We have argued that our results undermine Peña et al. and Endress and Bonatti's argument for the MOM hypothesis, and we therefore conclude that they have not demonstrated that rule-governed language-learning mechanisms are necessary for the extraction of structural information.

Now, mastering a language requires behaving (e.g., producing utterances) in accordance with non-adjacent dependencies, including long-distance dependencies. Common examples in the literature include the sort of dependencies that are required for maintaining agreement over lengthy center embeddings, as in ''The cats who the dog bites run.'' But the artificial language of Peña et al. and Endress and Bonatti, which we have adopted here for the purposes of direct comparison with their results, represents only a small subset of the full range of non-adjacent dependencies found in natural languages. One might therefore question the sense in which the studies reported here on structural rule learning in *artificial* language acquisition relate to structural rule learning in *natural* language acquisition. Illustrations abound, but simply consider parasynthesis, or infixation as cases where non-adjacent dependencies occur at the morphosyntactic level. Probably, parasynthesis, infixation, or agreement, are not straightforward natural language counterparts of the sort of non-adjacent

dependencies that define artificial rule words in Peña et al.'s and Endress and Bonatti's corpora. However, insofar as their results are thought to back up the MOM hypothesis, our modeling results may equally go beyond the idiosyncrasies of SRNs. Our work shows that a primitive, artificial statistical learning mechanism can learn linguistic preferences that appear to be governed by abstract, structural rules. There is no reason to think that the powerful statistical learning machinery of the human brain could not do the same.

Note

1. Endress and Bonatti hypothesize the existence of ''a general mechanism representing syllables in words as variables, capable of operating under a variety of input conditions. Such a mechanism would be able to extract relations between such variables within their respective units… The silences may act as 'markers' that define the units of an analysis. Such markers may be a prerequisite for dependencies between classes in speech to be analyzed, and this would explain [why] the mechanism for generalization seems to only work over an already segmented input'' (p. 291).

Acknowledgments

This research was supported by DGICYT Projects HUM2006-11603-C02-01 (Spanish Ministry of Science and Education and Feder Funds) and FFI2009-13416-C02-01 (Spanish Ministry of Science and Innovation) to AL and PC, and by Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia, through project 11944/PHCS/09 to PC. This paper extends a preliminary study that appeared in the *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. The authors contributed equally to both papers.

References

Altmann, G. T. (2002). Learning and development in neural networks – the importance of prior experience. *Cognition*, *85*(2), B43–B50.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, England: Oxford University Press.

Boden, M., & Wiles, J. (2000). Context-free and context-sensitive dynamics in recurrent neural networks. *Connection Science: Journal of Neural Computing, Artificial Intelligence & Cognitive Research*, *12*(3–4), 197–210.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2006). How to hit Scylla without avoiding Charybdis: Comment on Perruchet, Tyler, Galland, and Peereman (2004). *Journal of Experimental Psychology: General*, *135*(2), 314–321.

Chalup, S. K., & Blair, A. D. (2003). Incremental training of first order recurrent neural networks to predict a context-sensitive language. *Neural Networks*, *16*(7), 955–972.

Chomsky, N. (1980). *Rules and representations*. Oxford, England: Basil Blackwell.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science*, *23*, 417–437.

Christiansen, M., & Curtin, S. (1999). Transfer of learning: Rule acquisition or statistical learning? *Trends in Cognitive Science*, *3*(8), 289–290.

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

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

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

Gerken, L. A., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*(2), 183–206.

Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalisation in connectionist networks. *Language & Cognitive Processes*, *10*(6), 601–630.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*(2), 73–137.

Laakso, A., & Calvo, P. (2008). A connectionist simulation of structural rule learning in language acquisition. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 709–714). Austin, TX: Cognitive Science Society.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

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

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition (vol. 2: psychological and biological models)*. Cambridge, MA: MIT Press.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182.

Newport, E. L., & Aslin, R. N. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *BUCLD 24: Proceedings of the 24th Annual Boston University Conference on Language Development* 1–21. Boston: Cascalla Press.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225–237.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002a). Signal-driven computations in speech processing. *Science*, *298*(5593), 604–607.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002b). Signal-driven computations in speech processing (online supplement on materials and methods). *Science*, *298*(5593). Available at: http://www.sciencemag.org/cgi/content/full/1072901/DC1

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

Perruchet, P., Peereman, R., & Tyler, M. D. (2006). Do we need algebraic-like computations? A reply to Bonatti, Peña, Nespor, and Mehler (2006). *Journal of Experimental Psychology: General*, *135*(2), 322–326.

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, *34*(2), 255–285.

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), 573–583.
Perruchet, P., & Vinter, A. (1998). PARSER: A model of word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.
Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.
Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*(4), 463–490.
Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*, *45*(1), 45–94.
Redington, M., Chater, N., & Finch, S. P. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.
Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, *13*(9), 2093–2118.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926–1928.
Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory & Language*, *35*(4), 606–621.
Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.
Seidenberg, M. S., & Elman, J. L. (1999a). Do infants learn grammar with algebra or statistics? *Science*, *284*, 435–436.
Seidenberg, M. S., & Elman, J. L. (1999b). Networks are not 'hidden rules.'. *Trends in Cognitive Sciences*, *3*(8), 288–289.
Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop? *Science*, *298*, 553–554.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Appendix A: Supplemental information on methods

The words in the familiarization stream were randomly selected subject to the constraints that (a) a word of a given family cannot be immediately followed by another word of the same family; and (b) a word with a given intermediate syllable cannot be immediately followed by another word with the same intermediate syllable.

The online methods supplement to the original paper (Peña, Bonatti, Nespor, & Mehler, 2002b) states that rule (b) is in effect (''two words are not adjacent if they have the same intermediate syllable,'' p. 1). However, Fig. 1 of the original paper (Peña et al., 2002a) indicates that the familiarization stream 1 contains the string *puraki beliga tafodu pufoki talidu beraga*, which contains a violation of rule (b). Endress and Bonatti (2007) state that they use rule (b) (viz. ''Consecutive items could not belong to the same family, nor could they have the same middle syllable,'' in the description of the familiarization procedure for their Experiment 1, p. 255). Thus, for the purpose of the simulations, we made an attempt to constrain all training input according to rule (b).

However, there is a further issue, in that observing both of these constraints (a and b) makes it difficult to generate a sequence of exactly the required length. Depending upon how the initial words are randomly ordered, it can be impossible to place the last few words.

As an extreme example, imagine that the sequence started as follows (the superscripts indicate repetitions):

$$[pulikiberaga]^{100}[pufokibeliga]^{100}[purakibefoga]^{100}talidu$$

This sequence is only 601 words long, but it cannot be completed because all remaining words are from the same family. Of course, pseudorandomly selecting words tends to avoid such extreme examples. Nevertheless, a pseudorandom selection procedure remains susceptible to less extreme forms of the problem: The placement of words earlier in the sequence usually makes it impossible to place all words in the sequence while obeying constraints (a) and (b). When the word sequence is generated stochastically, it is highly unlikely for the problem to be this extreme. However, it is also unlikely that exactly 900 words can be used. In one trial run, for example, only 48 of 1,000 randomly generated sequences were able to use all 900 elements.

Perhaps of even greater concern, the sequences that do use all 900 elements tend to end with relatively long sequences of alternating word pairs. This is because the procedure for selecting words is essentially selection from a 900-element set without replacement. Thus, the 900th word of a 900-word sequence has 0 entropy—it is the only option left in the pool. The first 900-word sequence we generated by pseudorandom selection ended with 16 straight repetitions of ''*talidu*, *pufoki*.''

Appendix B: Descriptive and inferential statistics for Study 1

Mean cosine similarity values for each test item type after 50, 70, 100, and 200 epochs of training are shown in Table B1.

A within-subject ANOVA with test word type (four levels: class word, part word, rule word and word) as the independent variable and cosine similarity after 50 epochs of training as the dependent variable indicated significant differences among the mean cosine similarity values for test word types, $F(3,147) = 5860.157$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences (including the relevant comparisons:

Table B1
Mean cosine similarity values at selected epochs for 50 networks trained and tested with gaps

	W		RW		CW		PW		PW12		PW21	
Epochs	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	0.834	0.023	0.794	0.042	0.312	0.035	0.069	0.016	0.073	0.031	0.065	0.021
70	0.931	0.043	0.888	0.062	0.160	0.037	0.083	0.043	0.121	0.090	0.045	0.021
100	0.959	0.054	0.916	0.071	0.091	0.042	0.090	0.062	0.153	0.132	0.026	0.018
200	0.915	0.136	0.813	0.166	0.055	0.053	0.106	0.077	0.203	0.157	0.008	0.015

Note. CW, class words; PW, part words; PW12, part words of type 12; PW21, part words of type 21 RW, rule words; W, words.

words vs. rule words, rule words vs. class words, and class words vs. part words) are significant ($p < .001$ in all cases).

A within-subject ANOVA with test word type as the independent variable and cosine similarity after 70 epochs of training as the dependent variable indicated significant differences among test word types, $F(3,147) = 4030.477$, $p < .001$. Bonferroni-corrected multiple comparisons again indicated that all pairwise differences were significant ($p < .001$ in all cases).

A within-subject ANOVA with test word type as the independent variable and cosine similarity after 100 epochs of training as the dependent variable indicated significant differences among test word types, $F(3,147) = 3158.695$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences except the difference between part words and class words were significant ($p > .9$ for part words vs. class words, $p < .001$ in all other cases).

A within-subject ANOVA with test word type as the independent variable and cosine similarity after 100 epochs of training as the dependent variable indicated significant differences among test word types, $F(3,147) = 862.222$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences were significant ($p = .004$ for part words vs. class words, $p < .001$ in all other cases).

A within-subject ANOVA with test item type (five levels: class word, part word of type 12, part word of type 21, rule word and word) as the independent variable and cosine similarity after 50 epochs of training as the dependent variable indicated significant differences among the mean cosine similarity values for test item types, $F(4, 196) = 6060.826$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences are significant ($p < .001$) *except* the difference of 0.008 between part words of type12 and part words of type 21.

A within-subject ANOVA with test item type as the independent variable and cosine similarity after 70 epochs of training as the dependent variable indicated significant differences among the mean cosine similarity values for test item types, $F(4, 196) = 2756.253$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences are significant. Indeed, with the exception of the comparison between class words and part words of type 12 ($p = .013$), all of the differences are very highly significant ($p < .001$).

A within-subject ANOVA with test item type as the independent variable and cosine similarity after 100 epochs of training as the dependent variable indicated significant differences among the cosine similarity values for test item types, $F(4,196) = 1838.041$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences are significant. Indeed, with the exception of the comparison between class words and part words of type 12 ($p = .030$), all of the differences are very highly significant ($p < .001$).

A within-subject ANOVA with test item type as the independent variable and cosine similarity after 200 epochs of training as the dependent variable indicated significant differences among the cosine similarity values for test item types, $F(4,196) = 677.378$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that all pairwise differences are very highly significant ($p < .001$).

Table C1
Mean cosine similarity values at selected epochs for 50 networks trained and tested without gaps

	W		RW		CW		PW		PW12		PW21	
Epochs	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
50	0.785	0.070	0.523	0.129	0.229	0.058	0.556	0.005	0.568	0.008	0.545	0.007
200	0.642	0.214	0.446	0.156	0.109	0.075	0.513	0.051	0.521	0.072	0.505	0.064

Note. CW, class words; PW, part words; PW12, part words of type 12; PW21, part words of type 21 RW, rule words; W, words.

Appendix C: Descriptive and inferential statistics for Study 2

Mean cosine similarity values for each test item type after 50 and 200 epochs of training are shown in Table C1.

A within-subject ANOVA with test word type (four levels: class word, part word, rule word, and word) as the independent variable and cosine similarity after 50 epochs of training as the dependent variable indicates significant differences among the mean cosine similarity values for test word types, $F(3, 147) = 555.086$, $p < .001$. Bonferroni-corrected multiple comparisons indicate that the difference between part words and rule words is not significant ($p = .447$; note that the variance for rule words is relatively high), but all other pairwise differences are significant ($p < .001$).

A within-subject ANOVA with test word type (four levels: class word, part word, rule word and word) as the independent variable and cosine similarity after 200 epochs of training as the dependent variable indicates significant differences among the mean cosine similarity values for test word types, $F(3,147) = 147.441$, $p < .001$. Bonferroni-corrected multiple comparisons indicated that the difference between part words and rule words was significant ($p = .033$), as were all other pairwise differences ($p < .001$).

A within-subject ANOVA with test word type (five levels: class word, part word of type 12, part word of type 21, rule word, and word) as the independent variable and cosine similarity after 50 epochs of training as the dependent variable indicates significant differences among the mean cosine similarity values for test word types, $F(4,196) = 481.368$, $p < .001$. Bonferroni-corrected multiple comparisons indicate that neither the difference between part words of type 12 and rule words ($p = .185$) nor the difference between part words of type 21 and rule words ($p > .9$) is significant, while all other pairwise comparisons—including the comparison between part words of type 12 and part words of type 21—are very highly significant ($p < .001$). The fact that the difference between the two part word types (0.023) is nearly same as the difference between part words of type 21 and rule words (0.022)—and moreover the fact that the difference between part words of type 12 and rule words is even bigger (0.045)—may seem puzzling, in light of the fact that the difference between the two part word types is significant but the differences between either part word type and rule words are not. The reason for this apparent anomaly is that the variance for rule words is substantially higher than the variance for either type of part word.

A within-subject ANOVA with test word type (five levels: class word, part word of type 12, part word of type 21, rule word and word) as the independent variable and cosine similarity after 200 epochs of training as the dependent variable indicates significant differences among the mean cosine similarity values for test word types, $F(4,196) = 128.973$, $p < .001$. Bonferroni-corrected multiple comparisons indicate that the difference between part words of type 12 and part words of type 21 is not significant ($p > .9$), nor is the difference between part words of type 21 and rule words ($p = .118$). The difference between part words of type 12 and rule words is significant ($p = .046$), and all the other differences are highly significant ($p < .005$).