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.

A connectionist simulation of structural rule learning in language acquisition

Aarre Laakso (alaakso@indiana.edu)
Department of Psychological & Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Paco Calvo (fjcalvo@um.es)
Departamento de Filosofía, Universidad de Murcia
Murcia, 30100 Spain

Abstract

According to a dual-mechanism hypothesis, although statistical computations based on nonadjacent transitional probabilities may suffice for speech segmentation, an additional rule-following mechanism is required in order to extract structural information out of the linguistic stream. We present a neural network study that shows how statistics alone can support the discovery of structural regularities, beyond the segmentation of speech, disconfirming the dual-mechanism hypothesis.

Keywords: language acquisition; statistical learning; nonadjacent dependencies; neural networks.

Introduction

Language acquisition is a central component of human development. A key question is whether language can be acquired solely by domain-general (e.g., statistical) learning mechanisms or whether domain-specific (e.g., algebraic) learning mechanisms are required. Peña, Bonatti, Nespor, & Mehler (2002) reported experimental evidence from French-speaking adults that they argued shows that humans use both statistical learning (to segment speech) and algebraic computations (to induce structural regularities such as grammatical rules). Subsequently, Endress & Bonatti (2007) replicated and extended the Peña et al. results with Italian-speaking adults and attempted to model them using connectionist networks. Endress & Bonatti argued that their failure to model the experimental results with connectionist networks demonstrated that associative learning mechanisms were insufficient for language learning. In this paper, we report a set of connectionist simulations that *does* model the experimental results. We conclude that Peña et al. and Endress & Bonatti have not demonstrated that rule-governed structure learning mechanisms are necessary for language acquisition.

Peña et al.'s experiments

The experiments in question test adult speakers' ability to (1) segment speech based on non-adjacent dependencies, and (2) generalize beyond the familiarization corpus. The experiments were based on roughly the same method as Newport & Aslin (2000). The artificial language consists of "words" that have the form A_iXC_i, where A_i, X and C_i are syllables. The subscripts on A and C indicate that the nonadjacent syllables are matched, such that the transitional probability between an A_i and the following C_i is 1.0. There are three X syllables, so the transitional probabilities between an A_i and an intermediate X and between an X and the final C_i are each 0.33. There are three word classes, and no two adjacent words in the speech stream may be from the same class, so the transitional probability 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*. Ten-minute and two-minute familiarization streams were produced by concatenating tokens of these nine words, randomly selected subject to the constraints that (a) a word of a given family could not be immediately followed by another word of the same family, and (b) a word with a given intermediate syllable could not be immediately followed by another word with the same intermediate syllable. In human experiments, the streams were converted to synthesized speech.

In the experiments reported in Peña et al. (2002), participants were asked to choose, after familiarization, between pairs of stimuli that could belong to three kinds of test items: words, part words and rule words. (In their Experiment 1, subjects had to choose between a word and a part word, after having been familiarized for 10 minutes to a continuous stream. In Experiments 2 and 3, participants had to choose between a part word and a rule word, after 10 minutes of familiarization on either a continuous or a segmented stream, respectively.) The "words" were simply items of the form A_iXC_i, that is, words that had appeared in the familiarization stream. The "part words" were also items that had appeared in the familiarization stream, but ones that straddled a word boundary. These could be of two types: "type 12" part words consisted of items having the form C_iA_jX, whereas "type 21" part words consisted of items having the form XC_iA_j. The type 12 part words are *dubefo, dubeli, dubera, dupufo, dupuli, dupura, gapufo, gapuli, gapura, gatafo, gatali, gatara, kibefo, kibeli, kibera, kitafo, kitali* and *kitara*. The type 21 part words are *foduga, foduki, fogapu, fogata, fokidu, fokiga, lidube, lidupu, ligapu, ligata, likibe, likita, radube, radupu, ragapu, ragata, rakibe* and

rakita. The "rule words" have the form $A_iX'C_i$, where X′ indicates a syllable that had appeared in the speech stream but never in the middle of a word. The rule words are *beduga, bekiga, bepuga, betaga, pubeki, puduki, pugaki, putaki, tabedu, tagadu, takidu* and *tapudu*. These are called "rule words" because participants who learned rules of the form "if the first syllable is A_i, then the last syllable is C_i" should find them familiar.

Peña et al.'s Experiment 1 supports the hypothesis that statistics allow for speech segmentation (subjects preferred words over part words). They claim that their Experiment 2, however, shows that statistics are not sufficient to extract structural information in a continuous familiarization corpus (subjects preferred part words over rule words). In their Experiment 3, a "subliminal" 25ms pause was inserted between each pair of words; although participants reported no awareness of such gaps, their presence did affect the results. Specifically, Peña et al. found that, when participants were trained on a speech stream with gaps, the participants subsequently did prefer rule words to part words at test. Experiments 4 and 5 tested for preference between part words and rule words after familiarization on a continuous and a segmented stream for 30 and 2 minutes, respectively (see Table 1). In the first case, subjects preferred part words over rule words. In the second case, they preferred rule words over part words.

Table 1: Summary of Peña et al.'s experimental results (w = word; pw = part word; rw = rule word).

Exp.	Stream	Duration famil	Test choice
1	Continuous	10'	w over pw
2	Continuous	10'	no pref rw/pw
3	Segmented	10'	rw over pw
4	Continuous	30'	pw over rw
5	Segmented	2'	rw over pw

Peña et al. interpret the results of Table 1 as evidence for a dual-mechanism 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).

Endress & Bonatti's experiments

Endress and Bonatti (2007) go a step further and argue that subjects may not prefer rule-words themselves, but so-called "class words", which involve a higher level of abstraction. Class words have the form $A_iX'C_j$, that is, an A syllable from one class, followed by a syllable that had appeared in the speech stream but never in the middle of a word, followed by a C syllable from a different 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" (where the A class comprises syllables A_1, A_2 and A_3, and the C class comprises syllables C_1, C_2 and C_3). The class words are *beduki, bekidu, bepudu, bepuki, betadu, betaki, pubedu, pubega, puduga, pugadu, putadu, putaga, tabega, tabeki, tagaki, takiga, tapuga* and *tapuki*.

As Endress & Bonatti point out, the experimental results from Peña et al. (2002) highlight a negative correlation between structural generalization and familiarization duration. Likewise, in the case of class words, Endress & Bonatti assume that following algebraic computations results in a preference for generalization in the case of shorter familiarization durations, whereas a statistical mechanism should take longer to generalize. So, they predict that preference for class words will decrease for longer familiarization durations. The following table summarizes some of Endress & Bonatti's experimental results.

Table 2: Summary of Endress & Bonatti's results (cw = class word).

Exp.	Stream	Duration famil	Test choice
1	Segmented	10'	cw over pw
2	Continuous	10'	no pref cw/pw
3	Segmented	2'	cw over pw
4	Segmented	30'	no pref cw/pw
5	Segmented	60'	pw over cw
8	Segmented	2'	w over rw
12	Segmented	2'	rw over cw

Endress and Bonatti next report a set of studies with artificial neural networks that they claim shows that a Simple Recurrent Network, or SRN (Elman, 1990) cannot account for the preference for class words exhibited by humans in their experiments. In what follows, we report a set of SRN studies that *does* model the experimental results.

Study 1

The first simulation study was designed to find a set of network parameters that could learn the familiarization sequence quickly. Like Endress & Bonatti, we used an SRN. The syllables were coded as nine- or ten-bit pairwise orthonormal binary vectors (a "1-of-c" encoding). Networks trained without gaps in the input stream had nine input units. Those trained with gaps had ten input units, the tenth representing the gap. 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.

Endress & Bonatti do not report the activation functions or objective function used in their simulations. In our first set of simulations, we used the standard sigmoid activation function at both hidden-layer units and output-layer units, together with the sum-squared error objective function. In

an effort to follow Endress & Bonatti as closely as possible, we trained 20 different "subjects" (networks starting with different initial random weights) with five hidden units at each combination of the following learning parameters: epochs ∈ (10, 50, 90, 100, 500), learning rate ∈ (0.00001, 0.00005, 0.00009, 0.0001, 0.0005, 0.0009, 0.001, 0.005, 0.009, 0.01, 0.05, 0.09, 0.1, 0.5, 0.9) and momentum ∈ (0.1, 0.5, 0.9). Not one of these networks learned the problem well enough to get even a single output pattern correct within a tolerance of 0.2 (i.e., all units with target 0 having activations of 0.2 or less and the unit with target 1 having an activation of 0.8 or greater). In addition to the parameters we sampled, Endress & Bonatti also trained networks with five hidden units for 900, 1000 and 5000 epochs and networks with 27 hidden units on all combinations. It is possible that, had we tried networks with 27 hidden units or trained for a larger number of epochs, we would have found networks that could perform the task. However, we suspect that the problem lies elsewhere.

Study 2

There is a well-known issue with using sigmoid output units and the sum-squared error function to train networks on problems where the target patterns are mostly zeros. 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 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 usual 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. 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, we ran a second set of simulations using the softmax activation function at the output layer and the cross entropy objective function. For networks with nine input units (those trained without gaps in the input), we stopped training when networks got at least 33% of the training patterns right, because only ⅓ of the syllables are deterministically predictable (by the nonadjacent dependency). Networks with 27 hidden units reached this criterion in fewer than 300 epochs on average (N=20, M=259.3, SD=79.2). Even at 8000 epochs of training, networks with five hidden units had learned only about 10% of the patterns on average, and none of them had reached criterion (N=5, M=10.22, SD=4.69).

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=18). 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 then performed an ANOVA on the cosine values, with item type (training, part word type 12, part word type 21, rule word and class word) as a between subjects factor.

Results

The ANOVA showed a significant effect of item type (F(4,19)=96.014, p<0.001). Bonferroni-adjusted post-hoc comparisons showed that the differences between part words of type 12, part words of type 21 and rule words were not significant (p>0.05) but all other differences were significant (p<0.001). See Figure 1.

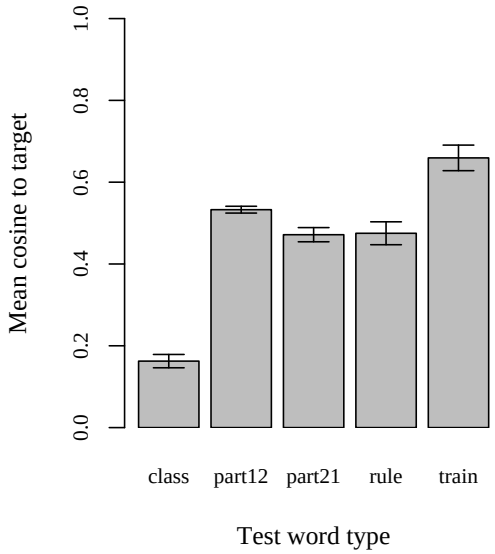

Figure 1: Mean cosine similarity between network outputs and target syllables for networks trained without gaps. Error bars in all figures show standard error.

Discussion

The results of our Study 2 accurately model the behavior of human participants in Experiments 1 and 2 of Peña et al. (2002), listed in Table 1. Specifically, Peña et al. found that human participants preferred training words to part words (their Experiment 1) but exhibited no preference for rule words over part words (their Experiment 2).

Study 3

In this study, we aimed to determine whether SRNs can exhibit a preference for rule words over part words, when

trained on a corpus that contains subliminal gaps (Peña et al. Experiment 3; see Table 1). For networks with ten input units (those trained with gaps in the input), we stopped training when they got at least 50% of the training patterns right, because the gaps, which followed every C syllable (¼ of the input patterns) were deterministically predictable from the preceding syllable, and ⅓ of the remaining syllables (the C syllables themselves) were deterministically predictable by the nonadjacent dependency. The networks trained with gaps learned the problem about twice as quickly as those without gaps, achieving criterion in about 150 epochs on average (N=20, M=153.5, SD=31.39).

The trained networks were tested in several ways. First, they were tested in exactly the same way as those in Study 2; in particular, no gaps were used before or within the test items. Second, they were tested with a gap at the beginning of every test item. Third, they were tested with a gap before every reliable A syllable. In this third case, at test, the training words, rule words and class words began with a gap, whereas the part words contained gaps between the first and second syllables (in the case of part words of type 12) or between the second and third syllables (in the case of part words of type 21). The third testing regime was designed to emulate the task reported by Peña et al. (2002) in their Footnote 27 (a control experiment intended to dismiss the possibility that a single statistical mechanism could be responsible for the preference of rule words found in their Experiment 3), and also simulated by Endress & Bonatti (2007) in a similar way. In test items with segmentation gaps, transitional probabilities for part words become higher than those for rule words, considering only adjacent transitional probabilities. However, once nonadjacent transitional probabilities are taken into account, the transitional probability of rule words becomes higher than that of part words. This means that participants in the control experiment may be computing statistical information about segmentation gaps. The prediction would be that they should favor rule words over part words, which is exactly what happens in Peña et al.'s control experiment.

Results

For networks tested without any gaps in the test items (Figure 2), the ANOVA showed a significant effect of item type (F(4,19)=370.49, p<0.001). Bonferroni-adjusted post-hoc comparisons showed that all differences were significant (p<0.001), except for the difference between class words and part words of type 12. For networks tested with a gap before every test item (Figure 3), the ANOVA again showed a significant effect of item type (F(4,19)=1123.9, p<0.001). Bonferroni-adjusted post-hoc comparisons showed that all differences were significant (p<0.001), except for the difference between class words and part words of type 21 (p=0.058). For networks tested with gaps within part words (Figure 4), the ANOVA again showed a significant effect of item type (F(4,19)=468.6, p<0.001). Bonferroni-adjusted post-hoc comparisons showed that all differences were significant (p<0.001).

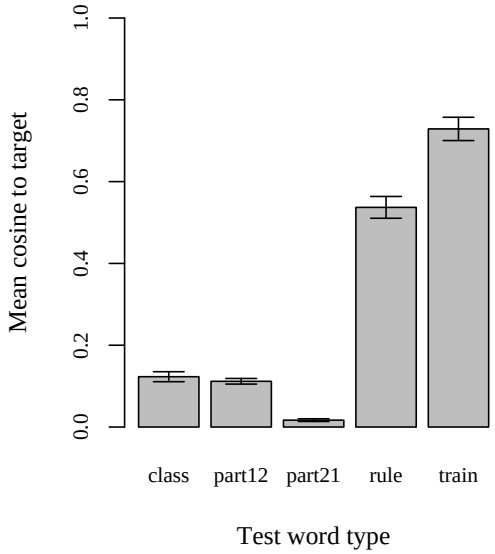

Figure 2: Results for networks trained with gaps and tested without gaps.

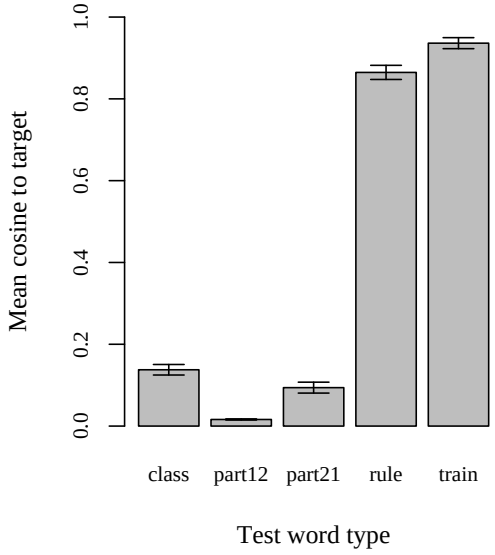

Figure 3: Results for networks trained with gaps and tested with a gap before every test item.

Discussion

The results of our Study 3 accurately model the behavior of human participants in Experiment 3 of Peña et al. (2002). In our simulations, 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. (2002). Thus, it is *not* necessary to suppose that non-statistical computations, "possibly of an algebraic or rule-governed nature…are

responsible for the observed behavior" (Peña et al., 2002, p. 606).

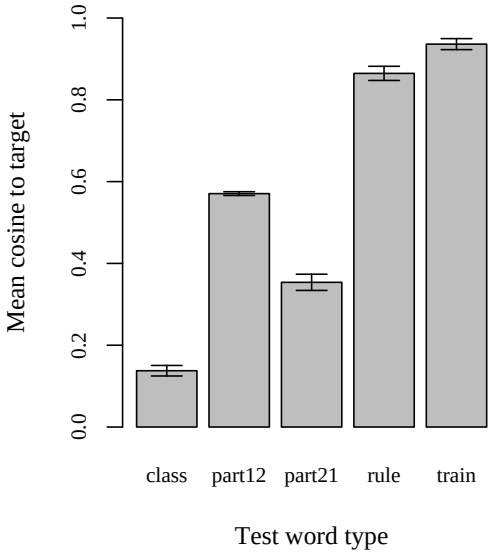

Figure 4: Results for networks trained with gaps and tested with gaps within part words and before other items.

Moreover, networks trained with gaps by our technique also exhibit a reliable preference for class words over part words of one or the other type. Networks tested without gaps prefer class words over part words of type 21, a result that Endress & Bonatti also reported for some of their networks trained with gaps and tested without. Endress & Bonatti dismissed this result because it predicted that, although human beings might prefer class words to part words of type 21, they would not prefer class words to part words of type 12, a result they had not observed in any of their experiments. However, our networks tested with gaps before test items prefer class words over part words of type 12, a reversal in the type of part words to which class words were preferred. The distinction between testing networks on part words that are preceded by gaps and testing networks on part words that are not preceded by gaps cannot be reproduced in the experimental procedure used with human beings – because the procedure involves comparing two different words presented separately, it is indeterminate whether the test words are "preceded by a gap" in the relevant sense. Indeed, it may be that some participants subconsciously align their calculations of transitional probabilities on the initial syllables of test words (which according to our simulations would lead to a preference for class words over part words of type 21), whereas others subconsciously align their calculations of transitional probabilities on the silences that precede test words (which according to our simulations would lead to a preference for class words over part words of type 12). Although Endress & Bonatti report no difference in the population mean responses to tests of class words versus the two types of part words, they do not report whether there are individual differences in preferences for class words over part words of type 12 versus part words of type 21. Finally, although networks tested with a gap before every test item (which clearly prefer class words to part words of type 12) do not prefer class words to part words of type 21 at the alpha=0.05 significance level, there is definitely a trend in that direction. It may be that networks trained somewhat less or more would exhibit a reliable preference for class words over both types of part words, a possibility that we explore in Study 4.

Study 4

Endress & Bonatti's (2007) results on segmented 2-minute familiarization streams (Experiments 3, 8 and 12 in Table 2) indicate preferences for words over rule words, for rule words over class words, and for class words over part words of types 12 and 21. To demonstrate that an SRN can model this pattern of preferences, we trained 20 networks with ten input units each for 30, 60, 90, 120, and 150 epochs of training, and tested them with a gap at the beginning of every test item. The goal was to determine, first, if class words are preferred over both types of part words, and, second, if the rank-order preference found by Endress & Bonatti (words > rule words > class words > part words) can be modeled statistically.

Results

The mean performance for each test item type is plotted as a function of the number of epochs of training in Figure 5. For networks trained for 120 epochs, an ANOVA showed a significant effect of item type ($F(4,19)=586.66$, $p<0.001$) and Bonferroni-adjusted post-hoc comparisons showed that all differences were significant ($p<0.001$), except the difference between training words and rule words ($p=0.18$).

Discussion

Overall, the results of our Study 4 do model the behavior of human participants in Experiments 3, 8 and 12 of Endress & Bonatti (2007). Initially, networks trained for 30 epochs exhibit a preference for class words over part words of both types, modeling the human behavior in their Experiment 3. However, no preference is observed between class words, rule words and training words. In that sense, the networks trained for only 30 epochs fail to match Endress & Bonatti's rank-order preference, because there are no differences between performance on class words, rule words and training words (Experiments 8 and 12). However, as the networks are probed after 60, 90, 120 and 150 epochs of training, performance on class words declines, while performance on training words improves more quickly than performance on rule words. Although the differences between training words and rule words in Figure 5 are not statistically significant, it is clear that the trend is toward better performance on training words than rule words.

Moreover, the results of our Study 3 demonstrate that, in networks that have learned to predict the training patterns, performance on rule words is reliably lower.

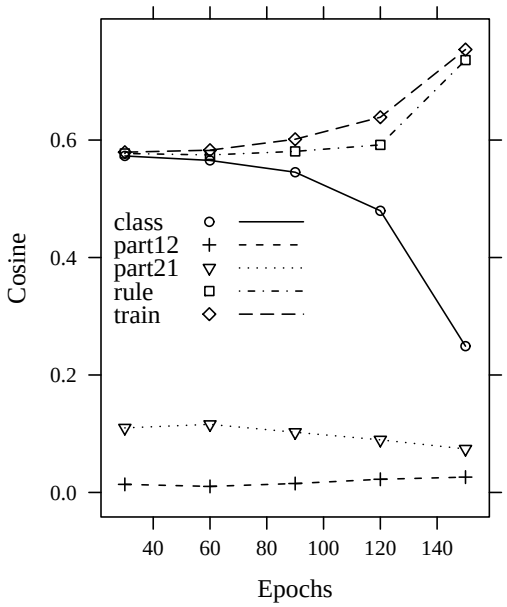

Figure 5: Results for networks trained with gaps for 30, 60, 90, 120 & 150 epochs, tested with gaps before items.

It may be argued nonetheless that, whereas in Experiments 3, 8 and 12 of Endress & Bonatti, subjects were always familiarized for 2 minutes, we have probed our networks at 30 epoch intervals between 30 and 150 epochs of training. In what sense, then, does Study 4 model the behavior of human participants? We chose to start with 30 epochs because the networks in our Study 3 trained for 150 epochs on average and the "short" streams in the human experiments were 2 minutes versus 10. There is no reason to expect, however, that there should be a proportional relation between epochs of training in artificial neural networks and familiarization duration with human subjects. In any case, the key point is that the networks do reproduce the preference for class-words over part words of both types after just 30 epochs of training, and that such a preference does not decay in subsequent epoch intervals as a result of a potential over-learning of the prediction task. That is to say, the networks retain the acquired knowledge of the structural regularities inherent to class words.

Second, the fact that the networks trained for only 30 epochs do not distinguish between class words, rule words and training words suggests perfect learning of the most abstract "rule", the one defining class words. The dual-mechanism hypothesis capitalizes on an observed negative correlation between the extraction of structural regularities and familiarization duration (Experiments 4 and 5, Table 1). 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 the rule (preference of class and rule words). However, the dual-mechanism hypothesis ignores the possibility that subliminal segmentation gaps can be exploited statistically, as the present results with SRNs illustrate.

Conclusions

According to a dual-mechanism hypothesis (Peña et al., 2002; Endress & Bonatti, 2007), 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. Our simulations show that a single statistical mechanism can account for the data that has been used to motivate the dual-mechanism hypothesis. We therefore conclude that Peña et al. and Endress & Bonatti have not demonstrated that rule-governed language-learning mechanisms are necessary for the extraction of structural information. In addition, we believe that these modeling results 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 that is the human brain could not do the same.

Acknowledgments

Preparation of this manuscript was supported by DGICYT Project HUM2006-11603-C02-01 (Spanish Ministry of Science and Education and Feder Funds) to A.L. and P.C.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

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.

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*. Boston: Cascalla Press.

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