On the Generalization of Non-Adjacent Dependencies: The Discrepancy between SRN Simulations and Human Behavior

Isao Inoue

1. Introduction

Recent studies in artificial language learning reveal that human language learners, both adults and infants, are well equipped with remarkable statistical computational power that enables them to segment words from seamlessly concatenated speech stream on the basis of conditional probabilities. The conditional probability is used as an objective method to assess the predictive power of one element with respect to another in sequential strings. When the occurrence of element *X* predicts the following occurrence of *Y* in the string, this predictive relationship can be considered to be the probability of the occurrence of *Y* given that *X* has previously occurred, which is written as P(*Y* | *X*), and is defined as:

(1) $P(Y \mid X) = P(X, Y) / P(X)$

where P(*X*) is the probability of *X*, and P(*X*, *Y*) is the joint probability of *X* and *Y*, indicating how often *X* and *Y* occur together. Because words are composed of specific sequence of syllables, the conditional probabilities between word-internal successive syllables will tend to be higher than those when syllable pairs in speech stream span word boundaries. If a language learner is sensitive to this statistic difference, there is a possibility that a learner could find word boundaries in speech stream given such distributional information. It has been recently shown by various artificial language learning experiments that human learners are capable of segmenting words by using only such statistical regularities among adjacent syllables (see Aslin, Saffran, and Newport (1998), Saffran (2001a), Saffran (2003), Saffran, Aslin, and Newport (1996), and Saffran, Newport, and Aslin (1996)). By 8 months, infants are found to be able to track conditional probabilities between adjacent syllables in continuous speech, using such information to identify words (Aslin, Saffran, and Newport (1998)).

Gómez and Gerken (1999) have reported that by 12 months infants are able to discriminate between grammatical and ungrammatical strings, based solely on conditional probabilities between adjacent words. In these experiments, infants showed remarkable ability to identify grammatical new strings, never encountered during familiarization phase. This shows that infants' statistical learning mechanisms can extract abstract distributional patterns which can be applied to judge novel strings.

However, the availability of such sensitivity to adjacent dependencies does not afford sufficient computational power to acquire natural languages, because remote distance dependencies between elements in strings are commonly observed in natural languages, as illustrated by various types of agreements among words in sentences. Gómez and Maye (2005) have observed that 18-month-old infants begin to exhibit such ability so that it takes about additional six months for infants to develop their ability to detect remote (or nonadjacent) dependencies in strings.

In artificial grammar learning paradigm, Gómez (2002) and Gómez and Maye (2005) have investigated how variability of the middle element *X* in the three-element string *AXB* affects the learnability of nonadjacent dependency relation between the elements *A* and *B*. It is found that high variability of the middle element gets adults and infants to detect nonadjacent dependencies. High variability of *X* means that the conditional probabilities between adjacent elements *AX* and *XB* are very low, in contrast to the nonadjacent *AB* pairing with the conditional probability of 1. Because only the nonadjacent *AB* pairings give us reliable statistical regularity to characterize stimulus strings under the high variability condition, learners are driven to focus on such structural regularity and disregard unstable adjacent dependencies involving the highly variable middle element *X*, despite the fact that computation of adjacent dependencies is much easier and acquired earlier.

In follow-up research on statistical learning of three-element strings of the form *AXB*, Onnis, Christiansen, Chater, and Gómez (2003), and Onnis, Monaghan, Christiansen, and Chater (2004) have found that zero-variability of the middle element *X* also facilitates the learning of nonadjacent *AB* pairs. Thus, only the middle level variability of *X* yields poor acquisition results so that the relationship between three variability conditions (zero, small, large) and the learning of nonadjacent dependencies is characterized as forming a U-shaped curve.

In this paper, I would like to study whether or not human learners' U-shaped performance on nonadjacent dependencies can be simulated by the Simple Recurrent Neural Network (SRN) (Elman (1990)), implemented by *Tlearn* software package

(Plunkett and Elman (1997)). We test whether SRNs can predict correctly the third nonadjacent element *B* when a novel item, not encountered during training phase, is used as a middle element *X* in *AXB*.

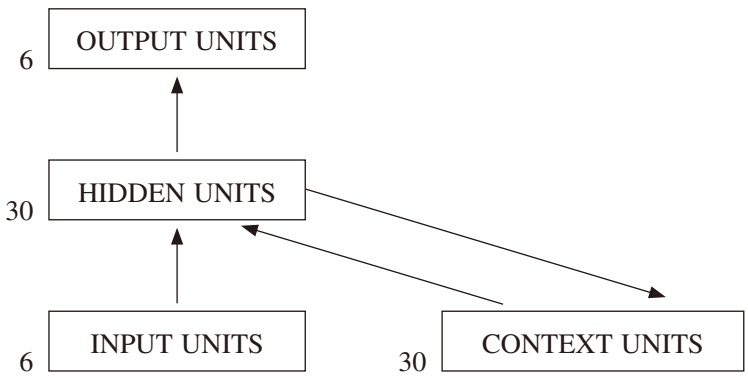
2. SRN Simulations

The SRN used in all our simulations has the following network architecture involving 6 input/output units and 30 hidden units which are recurrently connected, on a one-to-one basis, to 30 context units (see Figure 1). Context units function as dynamic memory, receiving the hidden node activation at a particular time step *t* and feeding it back to the same hidden node at the next time step *t*+1. Thus, hidden units can integrate the current input with their own previous processing result successively, obtaining computational power to detect nonadjacent dependency relationship between elements in strings.

In all our simulations the identical parameter settings are used: learning rate 0.1; momentum 0.3; initial weight randomization limit 0.1.

Since the present simulations are conducted to replicate human learning performance reported in Onnis et al. (2004), the three set-sizes of *X* (1, 2, 24) adopted in Onnis et al. (2004) are employed in the context of three nonadjacent pairings A_1__B_1, A_2__B_2, A_3__B_3. Each element in *AXB* strings is represented as a randomly selected distributed pattern of activity across a 6-bit vector:

Figure 1.



A_1	(0, 1, 1, 0, 0, 1)	B_1	(0, 0, 0, 1, 1, 0)
A_2	(1, 1, 0, 0, 0, 1)	B_2	(0, 1, 0, 1, 0, 0)
A_3	(1, 1, 0, 1, 1, 1)	B_3	(0, 1, 1, 1, 1, 0)
X_1	(0, 0, 0, 0, 1, 1)	X_2	(0, 0, 1, 0, 1, 0)
X_3	(1, 0, 1, 0, 1, 0)	X_4	(1, 1, 0, 1, 1, 0)
X_5	(0, 1, 0, 1, 0, 1)	X_6	(0, 0, 1, 1, 1, 0)
X_7	(1, 0, 1, 0, 1, 1)	X_8	(1, 1, 0, 0, 1, 1)
X_9	(0, 0, 1, 0, 0, 1)	X_{10}	(0, 0, 1, 0, 0, 0)
X_{11}	(1, 0, 0, 0, 0, 1)	X_{12}	(1, 1, 1, 1, 1, 0)
X_{13}	(1, 1, 1, 0, 0, 0)	X_{14}	(1, 0, 0, 1, 0, 0)
X_{15}	(0, 1, 1, 0, 1, 1)	X_{16}	(1, 1, 1, 0, 1, 0)
X_{17}	(1, 0, 1, 0, 0, 0)	X_{18}	(1, 1, 1, 1, 1, 1)
X_{19}	(0, 0, 1, 1, 1, 1)	X_{20}	(0, 1, 1, 1, 1, 1)
X_{21}	(1, 1, 0, 0, 1, 0)	X_{22}	(1, 1, 0, 1, 0, 0)
X_{23}	(1, 1, 0, 1, 0, 1)	X_{24}	(0, 0, 0, 0, 0, 1)

Because strings are separated by 750-ms pauses in Onnis et al. (2004), the following pause vector is inserted between *AXB* strings in our simulations:

pause　(0, 0, 0, 0, 0, 0)

72 *AXB* strings constitute one training set, which means one iteration of 3 (nonadjacent pairings) × 24 (*X* elements) strings in the case of set size 24, and 12 iterations of 3 (nonadjacent pairings) × 2 (*X* elements) strings in the case of set size 2, and 24 iterations for set size 1. Thus, 72 strings constitute one epoch of training, with the frequency of occurrence of nonadjacent pairings being held constant across the three variability conditions. For each of the three variability conditions, we conducted 12 simulations involving different initial startup configuration of the weights. After training SRNs for 200000 epochs, they were tested on strings involving novel *X* elements, not used in training phase. We adopted the rounding off criterion for the successful prediction of *B* elements. When all the six components of output activation vector are within 0.5 of the target values, i.e., $|\text{target} - \text{output}| \leqq 0.5$, the output should be considered correct. For example, if the target vector is (0, 1, 1, 1, 1, 0) and SRN gives the output vector (0.002, 1, 0.917, 0.998, 0.737, 0.004), then the absolute error is (0.002, 0, 0.083, 0.002, 0.263, 0.004). Because all the six components do not exceed the limit of 0.5, this output is considered correct. On the other hand, if the absolute error is (0.005, 0.001, 0.601, 0.002, 0.416, 0.006), this case should be considered incorrect, because the one component exceeds the 0.5 limit.

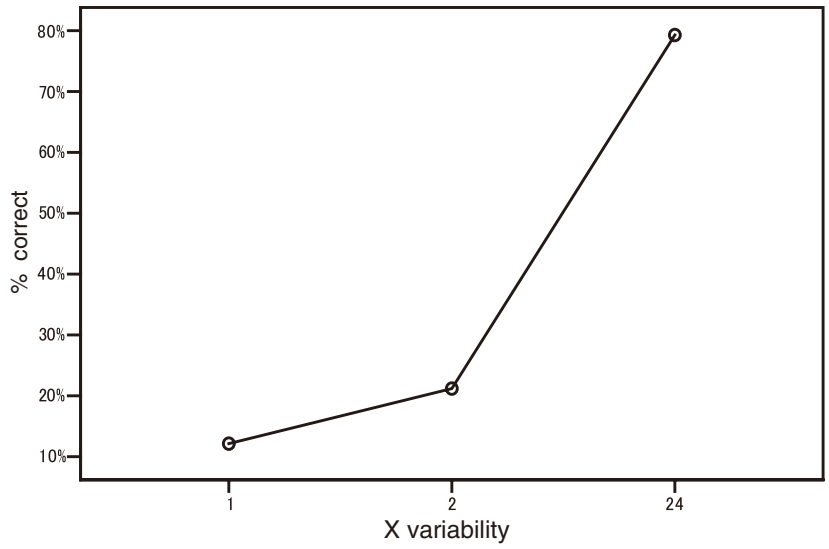
3. Analysis of SRN performance

We ran our simulations by using 12 SRNs with randomly assigned different initial state. Since the performance of each SRN was measured repeatedly under three variability conditions, a repeated-measures analysis of variance was conducted to evaluate the relationship between *X* variability and the predictability of the third element *B*. The independent variable included three levels of *X* variability factor (1 vs. 2 vs. 24) and the dependent variable was the correct prediction of *B* expressed as a percentage. The means and standard deviations for correct prediction scores are represented in Table 1. The results of the ANOVA indicated a significant *X* variability effect, Wilks' Lambda = 0.085, $F(2, 10)=0.915$, $p<0.001$.We conducted three pairwise comparisons among the means for 1, 2, 24 *X* variability. All the pairwise comparisons were significant controlling for familywise error rate at the 0.05 level, using Bonferroni procedure. The percentage of correct prediction increases in parallel with *X* variability, as shown in Figure 2.

Table 1

X variability	Mean	Std. Deviation	N
1	12.0370	6.22070	12
2	21.0648	11.18713	12
24	79.2824	21.90796	12

Figure 2. Generalisation under variability



This consistent upward trend presents a marked contrast to the human learners' U-shaped learning curve observed in Onnis et al. (2004). While SRNs gave extremely poor performance on generalization under zero variability (*X* set size 1) condition, human learners achieved the very high score, reaching almost 90% level, much better than approximately 65% correct response under small variability (*X* set size 2) condition. Human learners' performance under large variability (*X* set size 24) condition was relatively high (roughly 75% correct), which was successfully modeled by our SRN simulation.

Gómez (2002) suggests that when human learners attempt to process strings, they try to analyze sequential structure by computing conditional probabilities between adjacent elements as the default, and only when the computation of adjacent dependencies gives no reliable information about the string structure, they turn to the computation of nonadjacent dependency even though its computation load is much heavier than that of adjacent dependency. The computation of conditional probabilities involving highly variable middle elements gives us no reliable information about string structure. This explains why the large variability leads human learners to switch their focus to nonadjacent dependency and to disregard the highly variable middle elements. On the other hand, the conditional probabilities of adjacent elements under small variability condition provides relatively reliable information about the sequential structure of strings, and hence no further computation is necessary to get additional information about the sequential structure. This is the reason why small variability yields poor performance on the acquisition of nonadjacent dependencies.

When we process input strings (2a, b, c) with zero variability in middle position, their sequential structure can be gained either by computing adjacent dependency of AX_i and X_iB where the middle element X_i is fixed or by computing nonadjacent dependency involving three different *AB* pairings.

(2)　a.　$A_1\ X_i\ B_1$
　　b.　$A_2\ X_i\ B_2$
　　c.　$A_3\ X_i\ B_3$

Onnis et al. (2003) argue that because the middle element is always the same, the former approach produces low information gain with respect to the sequential structure of *AXB* strings and hence, learners are driven to adopt the latter approach, seeking another informative source of sequential information about input strings. The high conditional probabilities of nonadjacent *AB* pairings give us crucial information

about which specific *B* element comes in the third position by knowing the first element *A*. When human learners focus on nonadjacent dependency and disregard the middle element X_i as in the case of the latter approach, the acquired structural regularities can be represented as (3), where the middle position functions as a placeholder which accepts any *X*:

(3) a. $A_1 \_\_ B_1$
b. $A_2 \_\_ B_2$
c. $A_3 \_\_ B_3$

If (3a, b, c) are acquired under zero variability, human learners can discern correct *AB* pairings in strings involving not only X_i but also other new middle elements, giving good performance on generalization, as observed in Onnis et al. (2004).

Now, return to the performance of SRNs under zero variability. We trained our SRNs by using strings of the form (2a, b, c) as training data. After training, we evaluated the performance of SRNs on training data and found that SRNs have completely learned the patterns in the training set. They correctly predicted the occurrence of X_i at an average rate of 99.42% when they receive *A* (namely, it can be A_1, A_2 or A_3) as an initial input, and the correct prediction of the third element on receiving the middle element X_i as input also reached the high average percentage of 96.18. The highly successful learning of the training data exhibited clear contrast to the low success rate of 12.037% on generalization, as shown in Table 1. This means that the presence of the specific middle element X_i plays a crucial role in the correct prediction of the third element by SRNs trained under zero variability, so that the SRNs fail to model the middle element as a placeholder as represented in (3).

References

Aslin, R. N., J. R. Saffran, and E. L. Newport (1998) "Computation of Conditional Probability Statistics by 8-Month-Old Infants," *Psychological Science,* 9, 321–324.

Elman, J. L. (1990) "Finding Structure in Time," *Cognitive Science* 14, 179–211.

Gómez, R. L. (2002) "Variability and Detection of Invariant Structure," *Psychological Science* 13, 431–436.

Gómez, R. L. (in press) "Dynamically Guided Learning," in Y. Munakata and M. Johnson (eds.) *Attentional and Performance XXI: Processes of Change in Brain and Cognitive Development*, Oxford, Oxford University Press.

Gómez, R. L. and L. Gerken (1999) "Artificial Grammar Learning by 1-Year-Olds Leads to Specific and Abstract Knowledge," *Cognition* 70, 109–135.

Gómez, R. L. and L. Lakusta (2004) "A First Step in Form-Based Category Abstraction by 12-Month-Old Infants," *Developmental Science* 7, 567–580.

Gómez, R. L. and J. Maye (2005) “The Developmental Trajectory of Nonadjacent Dependency Learning,” *Infancy* 7 (2), 183–206.

Mintz, T. H. (2002) “Category Induction from Distributional Cues in an Artificial Language,” *Memory & Cognition* 30, 678–686.

Mintz, T. H. (2003) “Frequent Frames as a Cue for Grammatical Categories in Child Directed Speech,” *Cognition* 90, 91–117.

Mintz, T. H., E. L. Newport, and T. G. Bever (2002) “The Distributional Structure of Grammatical Categories in Speech to Young Children,” *Cognitive Science* 26, 393–424.

Newport, E. and R. Aslin (2000) “Innately Constrained Learning: Blending Old and New Approaches to Language Acquisition,” in S. C. Howell, S. A. Fish, and T. Keith-Lucas (eds.) *Proceedings of the 24th Annual Boston University Conference on Language Development*, pp. 1–21, Cascadilla Press, Somerville, Massachusetts.

Newport, E. L. and R. N. Aslin (2004) “Learning at a Distance I. Statistical Learning of Non-Adjacent Dependencies,” *Cognitive Psychology*, 48, 127–162.

Onnis, L., M. H. Christiansen, N. Chater, and R. Gómez (2003) “Reduction of Uncertainty in Human Sequential Learning: Evidence from Artificial Language Learning,” *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 886–891, Lawrence Erlbaum, Mahwah, New Jersey.

Onnis, L., P. Monaghan, M. H. Christiansen, and N. Chater (2004) “Variability Is the Spice of Learning, and a Crucial Ingredient for Detecting and Generalizing in Nonadjacent Dependencies,” *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, New Jersey.

Peña, M., L. Bonatti, M. Nespor, and J. Mehler (2002) “Signal-Driven Computations in Speech Processing,” *Science*, 298, 604–607.

Plunkett, K. and J. L. Elman (1997) *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations*, The MIT Press, Cambridge, Massachusetts.

Saffran, J. R. (2001a) “Words in a Sea of Sounds: The Output of Infant Statistical Learning,” *Cognition* 81, 149–169.

Saffran, J. R. (2001b) “The Use of Predictive Dependencies in Language Learning,” *Journal of Memory and Language* 44, 493–515.

Saffran, J. R. (2003) “Statistical Language Learning: Mechanisms and Constraints,” *Current Directions in Psychological Science* 12, 110–114.

Saffran, J. R., R. N. Aslin, and E. L. Newport (1996) “Statistical Learning by 8-Month-Old Infants,” *Science* 274, 1926–1928.

Saffran, J. R., E. L. Newport, and R. N. Aslin (1996) “Word Segmentation: The Role of Distributional Cues,” *Journal of Memory and Language* 35, 606–621.

Santelmann, L. M. and P. W. Jusczyk (1998) “Sensitivity to Discontinuous Dependencies in Language Learners: Evidence for Limitations in Processing Space,” *Cognition*, 69, 105–134.

Spiegel, R. and I. P. L. McLaren (2001) “Human Sequence Learning: Can Associations

Explain Everything?" in J.D. Moore and K. Stenning (eds.) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp.976–981, Lawrence Erlbaum, Mahwah, New Jersey.