Background

Almost since its inception, the field of linguistics has been entrenched in debate about the nature of the linguistic knowledge, and more specifically syntax, which human beings almost universally come to possess in one way or another. The most well-known theory which accounts for this knowledge states that all human languages are built from a universal grammar (UG), which is basically an innate set of rules which constrain the allowable sentence structures in a language (Chomsky, 2002).How humans come to acquire this very specific rule set is still an open question. One of the main arguments for this view has come to be known as the "poverty of the stimulus" argument, which posits that the number of potential grammars which a human learner could generate from the input which they receive is exceedingly large, yet children are able to ignore most of these dead-end options and focus in on the actual grammar they are learning within a few short years (Laurence, 2001). The "innate" hypothesis resolves this dilemma; if the basic rules are already contained in the child's head, learning a language becomes a relatively simple matter of tuning that machinery in to the specifics of the language being learned. Under this hypothesis, the existence of syntactic categories, movement rules, and other syntactic attributes is mostly inherent, making the task of learning to use them realistic from a child's perspective.

A competing theory is that syntax is to a greater or lesser extent emergent, meaning that syntactic categories and rules develop empirically over time as a result of outside influences. This theory states that, contrary to the poverty of the stimulus argument, there is sufficient information in children's input to derive such categories and attributes empirically, if only one knows where to look (Seidenberg, 2002). In this way, the acquisition of complex syntactic knowledge is accounted for without recourse to innate mechanisms. Because this hypothesis involves learning, it is still constrained in some ways; not by innate syntactic knowledge, but rather by the computational constraints of the learning process. Such constraints would be dependent on the type of learning involved. Since much of syntax acquisition is assumed to be unconscious, this learning is most likely implicit.

Implicit learning is the complement to explicit learning, that is, when instructions and/or feedback are given to an agent that is trying to learn a task. Contrarily, implicit learning occurs in the absence of overt guidelines, and is characterized by being an unconscious process which results in abstract, tacit knowledge about the surrounding environment (Reber, 1989). Much of language development seems to be implicit; children quickly learn to use language but are typically unable to articulate exactly how this knowledge operates. Research with implicit learning of artificial grammars supports this; both adults and children are able to make accurate grammatical judgments, but when questioned cannot give much explicit reasoning for their choices other than “it sounds right/wrong” and other similarly vague responses (Reber, 1967). Even more revealing, children with specific language impairment perform poorly on implicit learning tasks (Evans, 2009), further corroborating evidence for a link between implicit learning and language. Given the role played by implicit mechanisms in pattern learning tasks, most notably learning artificial grammars, general implicit learning mechanisms are likely to be a key component of language development (Reber, 1989).

One particular type of learning which is often considered to be implicit is statistical learning. Subjects in statistical learning tasks have been shown to be sensitive to small changes in statistical cues, even while being consciously unaware of any statistical tracking (Vouloumanos, 2009), qualifying their knowledge as implicit. There are a number of interesting statistical attributes to human language, including distributional information (Newport, 2000). One particularly salient statistical dependency is the conditional probability of one element occurring after another element (Saffran, 2001). These conditional probabilities are sometimes referred to as transitional probabilities because they can be used to delineate boundaries between larger components (Kam, 2009). This type of statistical cue can theoretically operate over any type of sequence, from syllables to words to syntactic phrases, and provides a very rich source of information from linguistic input that would otherwise be quite sparse (Newport, 2000).

One interesting aspect of research using transitional probabilities is “probability matching.” Probability matching is the tendency for participants in studies using transitional probabilities to produce or interpret structures or elements based on the distribution of their transitional probabilities (Kam, 2005). This leads to more common forms being interspersed with rarer forms with about the same frequencies as the input from which those forms are heard. In order to probability match, subjects must perform the relevant statistical computations over the given input. Thus, the ability to probability match is a convenient way to ascertain whether or not subjects are able to track statistics over particular linguistic features.

Adults, children, and even animals are able to segment a continuous speech stream into discrete words based only on the transitional probabilities between the syllable components of the words (Newport, 2000; Saffran & Wilson, 2003; Xie, 2012). Components with high transitional probabilities tend to belong to the same word, whereas components with low transitional probabilities usually indicate word boundaries. Equivalently, adults and children are able to group words together into rough syntactic categories based on their transitional probabilities (Williams, 2010); in this case, words with high transitional probabilities tend to belong to different categories while words with low transitional probabilities are more likely to belong to the same category because they have complimentary distributions within the sentence. This type of learning is not unique to language; adults, children, and many animals are able to differentiate between tone sequences using only their transitional probabilities (Saffran, 1999). Even other modalities can use this computational resource; both the visual system and the visuo-motor system can perform the same types of statistical categorization (Saffran, 2002).

Proposal for a Synthesized Linguistic Statistical Learning Architecture

Given the ubiquitous nature of statistical cues in language, it seems likely that statistical learning mechanisms are utilized at many levels of linguistic abstraction simultaneously (Saffran, 2003). If the output from one system can form the input to another, more abstract system, statistical learning mechanisms would potentially allow the learner to formulate grammatical sentences without ever needing to rely on explicit grammatical rules (an idea hinted at in Saffran, 2003). If each system operates in parallel, more abstract levels will initially have very little impact on overall performance because the input from lower layers is not coherent enough to make useful distinctions. If output is constrained to the most coherent form available, more abstract layers will be ignored until they become relevant, due to the lack of sufficiently detailed input. This predicts that the children’s initial output will be very rudimentary and will slowly increase in complexity as useful distinctions are found (leading from syllable discrimination to word segmentation, for example). Additionally, if feedback from more abstract layers is able to affect less abstract layers recursively, then as more abstract levels of language become relevant and start to be employed in production, lower levels would be predicted to fluctuate from their previous functionality.

In order to test this hypothesized architecture, it must first be shown that statistical learning mechanisms are capable of operating over highly abstract syntactic features like phrase categories. If this top-level ability exists, then further experimentation to assess the existence and operation of linked statistical learning mechanisms would be justified. The rest of this paper addresses the existence of highly abstract statistical learning by simulating the results of a recent study which examined subjects' ability to track transitional probabilities between phrase categories. This is done with the intention of providing evidence that highly abstract statistical information can be learned and represented through weighted connections, and that such information can be calculated from arbitrarily represented inputs, both of which are necessary conditions for a connected statistical learning architecture like the one proposed above.

Recent Work

A study carried out by Carla Hudson Kam (Kam, 2009) assessed the ability of human adults to compute and apply transitional probabilities between the syntactic phrases Subject, Object, and Verb. The experiment participants were exposed to an artificial grammar using realistic-sounding words arranged into sentences in varying word orders. Subjects in each condition were presented with a mix of transitive and intransitive sentences, with the experimental variable being the word order of the sentences. Roughly half of the sentences were transitive, the other half intransitive. The word orders for the intransitive sentences were the same for each condition, 60% of the sentences were VS (Verb followed by Subject) and 40% were SV. The probability of each noun occurring in each position was balance so that no noun occurred more often in Subject rather than Object position, and vice-versa. This set-up allowed the transitional probabilities within phrases to be manipulated explicitly by the experimenters. After accounting for other linguistic correlations, the only clue to the proper word order was the probability of V following S compared to S following V. The transitive sentences were formed into three conditions; one condition heard SOV 40% of the time (referred to as the SOV condition), the next heard SVO 40% of the time (SVO condition), and the last heard SOV 20% of the time and SVO 20% of the time (SOV & SVO condition). All three conditions heard VSO 60% of the time. After exposure, participants were given 3 tests; a vocabulary test to ensure that they had learned enough words to be able to use the syntax appropriately, a forced-choice grammar task to ensure that they understand the grammar correctly, and the experimental task, a production task which required them to produce sentences to describe particular events using vocabulary that they had learned.

The results of the study were quite revealing (see Figure 1). For the intransitive sentences, participants matched the transitional probabilities very closely, with almost 60% of produced sentences being in VS order while roughly 40% were in SV order. The results for the transitive conditions were much messier because of L1 interference (knowledge of a first language interfering with production of a second). There was a universal tendency to use SVO word order; since all subjects were monolingual English speakers, this tendency can be attributed to L1 interference. There is still a tendency to probability match for these conditions, though the interference makes it tough to distinguish. The presence of L1 interference was not actually surprising and actually had the benefit of providing solid evidence that whatever method the participants were using to perform the task was applicable to linguistics, a common criticism of statistical learning studies.

Simulation

One fruitful technique for modeling implicit sequential learning paradigms involves the use of simple recurrent neural networks (SRNs) (Chang, 2006). SRNs consist of an input layer, a hidden layer, a context layer, and an output layer (see Figure 2). The role of the context layer is to “remember”, or entrain, previous inputs, which provides a very simple method for the network to analyze sequences and recall the patterns inherent in them. The context layer has the same number of nodes as the hidden layer and is initially set either to zero or to a static value (such as 0.5). After the initial input is received and the initial output calculated, the activation values of the hidden layer are copied onto the context layer. The hidden layer receives inputs from both the input layer and the context layer, which ensures that the previous input will have an effect on the next input, allowing the network to incorporate temporal sequences into its output, which is essential for processing sequences correctly. These networks have been taught to recognize grammatical sentence structures and tested on their predictive performance about grammaticality judgments for novel sentence (Williams, 2010), and have been shown to degrade in a manner similar to human memory (Cleermans, 1991).

Given that statistical learning requires inputs to be parsed sequentially in order to establish estimated transitional probabilities between elements, it is reasonable to adapt the use of SRNs to this task. By constructing a nested set of these SRNs, with each layer taking as its inputs the outputs of a previous layer and the outputs of all layers converging to produce a single output, the earlier proposed architecture of nested layers of statistical learning mechanisms working in parallel can be modeled. Such a model is very difficult to construct from scratch because each component must function appropriately in order for the model to be valid and malfunctions in any component will skew the final output. However, by assuming that each component is essentially modular, only being affected by the rest of the system through its inputs and only affecting the system through its outputs, than any part of the overall network can be isolated and studied under the assumption that the components before it will provide it with the correct input. Such modular models have already been successfully modeled for various linguistic components (Williams, 2010), so the same assumption of modularity will be applied to the SRN adapted to simulate the results of the Hudson Kam study.

Since the variable of interest in the study was the transitional probabilities between phrase categories, the input to the modified SRN was designed to represent the presence or absence of each category in the input sentence. This follows from the assumption of modularity; a separate system would be responsible for identifying the various syntactic categories in the input and would provide this information to the modified SRN. In turn, the output of the modified SRN would be a word order, which would become a template around which the sentence would be constructed. To approximate this connection, a simple binary coding system was used in which each category was represented by a different input value, which was set to 1 if the category was present in the sentence and to 0 if the category was not present. Since only basic transitive and intransitive sentences were used, the input consisted of three values representing Subject (S), Object (O) and Verb (V) respectively, and only two different combinations of input values were possible. The output also consisted of three values, each representing the same categories. After the initial input was given, the output was run through a decision function which selected a category. The decision function generated a random value between zero and one and selected the output by comparing the random value to each output value in turn, with each output being added to the output preceding it before being compared to the random value, which ensures that the value of each output becomes the probability with which it will be chosen. The next input in the sequence was based of the chosen output category, which was set to 1 while the other two were set to 0. The process was continued until the network had selected enough categories to match the number of categories present in the original input. If the network is sensitive to transitional probabilities, then the number of different sentence orders generated would be expected to match the transitional probabilities of those sentence orders.

This simulation was built around a simple artificial neural net, which was intentionally designed to resemble the actual system of connections that exists between neurons in a real brain. However, this model of the neuron is a vast oversimplification, and actual synaptic interactions are much more complicated. Furthermore, there is no account made for L1 interference, and the network operates on theoretical syntactic categories instead of taking input directly from a speech stream. There are also no analogs for the grammar and vocabulary tasks given in the original experiment. On the one hand, this makes the simulation more artificial because the inputs that it assumes and the outputs it produces are only indirectly observable in the results from the human subjects. On the other hand, the simplicity of the simulation allows for the variable of interest (the transitional probabilities) to be assessed directly, which helps ensure that the results will not be confounded by outside variables. The results of this simulation will hopefully be able to shed some light on the validity of the assumption that statistical learning mechanisms can be functionally modular, and should be a valid test of whether transitional probabilities can be tracked using simple connectionist systems, but any further generalizations must be curtailed, as the validity of this simulation drops off sharply when applying the results to domains beyond those specifically stated.

Experiment

The basic experimental set-up consisted of 30 simulated subjects, each modeled by a separate modified SRN, being exposed to a set number of exposure sentences, then producing a set of 24 word orders. Since the exposure rates given by Hudson Kam were timed, the number of exposure sentences was approximated to 900 trials, assuming that each sentence in the original exposure took about 10 seconds (150 sentences per day + 2 half days = 900). The learning rates of each subject were random values between 0 and 1, the momentums will also be random values between 0 and .5 and all the other SRN parameters will be held equal. The networks each consist of 3 input nodes, 3 output nodes, 12 hidden nodes, and 12 context nodes, and were based off of the architecture presented in (Freeman, 1994). For each exposure sentence, the network was presented with the appropriate input, and then the weights were updated through backpropagation by using the word order of the exposure sentence as the desired outputs. This was done sequentially; once the first category had been backpropagated, it was then introduced as the input so the network could learn the second category appropriately, and so on.

For the production task, each subject was presented with the appropriate starting input, with the output run through the decision function becoming the next input. The final result was a string representing a particular category order, which was the sequence in which the simulated subject would have ordered the vocabulary items in the real experiment. This process was repeated 24 times per subject, as was done in the original experiment. Finally, the total number of productions of each category order was averaged, yielding the probabilities of each category order being chosen as the final results. This was done for 7 different conditions: (1) SVO without intransitives, (2) SOV without intransitives, (3) SVO and SOV without intransitives, (4) intransitives without transitive sentences, (5) regular SVO, (6) regular SOV, and (7) regular SVO and SOV.

Results

The first 4 conditions were intended to examine the simulated behavior without the potential confound of having both transitive and intransitive sentences being learned at the same time. Figure3 shows the results for conditions 1-3. There are a number of responses which did not fall into any category for all three conditions; this is not unexpected, and is a check on the validity of the simulation. The stochastic decision function can potentially choose incorrect category orders as well as fail to select any category over another for a particular position, simulating the human tendency towards fallibility. The presence of non-grammatical forms was consistent throughout all of the conditions. Condition 1 (SOV) showed fairly close matching (52%/27%), approaching the 60%/40% split present in the input, thus demonstrating basic probability matching. Condition 2 (SVO) had equivalent performance (48%/26%). Condition 3 (SVO and SOV) had similar performance (46%/14%/15%), with the VSO form being under-matched while the other two forms approached the 20% levels present in the input. Condition 4, which looked at intransitive sentences only (see Figure 4), showed poorer performance, but still moved toward the right probabilities (46%/20%). Conditions 5-7 (see figure 5) included both transitive and intransitive sentences. Performance was poorer overall, but still tended towards probability matching. Conditions 5 (40%/22% of the transitive sentences) and 6 (40%/20% of the transitive sentences) were roughly equal in performance, with the transitive and intransitive forms performing equally. Condition 7 shows slightly better performance, though the rarer forms (SVO and SOV) are under-matched (46%/8%/10% for transitive sentences, 42%/26% for intransitive sentences).

Discussion

The intransitive results (from Condition 4 especially, but from all the conditions to an extent) seem to coincide quite well with the results from the Hudson-Cam experiment after accounting for the ungrammatical forms (see Figure 4). This is the most critical result, as the intransitive matching was the aspect of the original experiment which best demonstrated probability matching. The other results are not really comparable to the results of the original study because of the absence of L1 interference, but they do show a marked tendency to probability match. Discounting the presence of L1 interference, the results do mirror the general trend of the original results.

The model underlying this simulation was kept simple to ensure that the results would represent the tracking of statistical probabilities as purely as possible. Many additions and changes could have been made to make the system more realistic. A mechanism mimicking human subjects' conscious ability to avoid obviously ungrammatical structures (e.g, category orders with repeated categories, such as VVV or OVO) would improve performance. The input and output could have been coded to represent individual words, making the function of the network be to reorder the words into the appropriate category order, making the system more valid (since it could use the exact same inputs as the actual human subjects). The system could also have been pre-trained on SVO pattern sentences to represent L1 influence. All of these options were eschewed in favor of simplicity of design, but given the robustness of the basic results seen in this experiment, it would likely be worthwhile to revisit this model and see how implementing these changes effects the results.

Conclusions

Although the simulation here presented here is very basic, it demonstrates that modular connective systems are capable of tracking and matching transitional probabilities. Because the inputs and outputs to this simulation were simple binary, they could theoretically represent any aspect of syntax, which would allow this model to be adapted to any level of syntactic abstraction if coded appropriately, from category distinction to syllable segmentation. By linking several such networks together, with feedback loops from higher levels to lower levels, a rudimentary mechanism for producing syntactically grammatical structures which uses statistical learning over exposure sentences emerges. The human brain certainly has the machinery to implement this type of statistical learning mechanism and has been reliably shown to be sensitive to statistical cues in linguistic input, even at very young ages. The most logical step for future research along these lines would be to begin to link several networks together and explore the effects of feedback between them.

References

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review*, *113*(2), 234.

Chomsky, N. (2002). *Syntactic structures*. de Gruyter Mouton.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language and Hearing Research, 52*(2), 321.

Freeman, J.A. (1994) Simulating Neural Networks with Mathematica. Reading, MA: Addison-Wesley.

Kam, C. L. H. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development*, *5*(2), 115-145.

Kam, C. L. H., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151-195.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British journal for the philosophy of science*, *52*(2), 217-276.

Newport, E. L., & Aslin, R. N. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In *Proceedings of the 24th Annual Boston University conference on language development* (Vol. 1).

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, *6*(6), 855-863.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, *118*(3), 219.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172-196.

Saffran, J. R., & Wilson, D. P. (2003). From Syllables to Syntax: Multilevel Statistical Learning by 12 Month-Old Infants. *Infancy*, *4*(2), 273-284.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current directions in psychological science*, *12*(4), 110-114.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R. (2001). The use of predictive dependencies in language learning.*Journal of Memory and Language*, *44*(4), 493-515.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop?. *Science*, *298*(5593), 553-554.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729-742.

Williams, J. N. (2010). Initial Incidental Acquisition of Word Order Regularities: Is It Just Sequence Learning?. *Language Learning*, *60*, 221-244.

Xie, Y. (2012). Transitional Probability and Word Segmentation. *International Journal of English Linguistics*, *2*(6), p27.

Figures

Figure 1

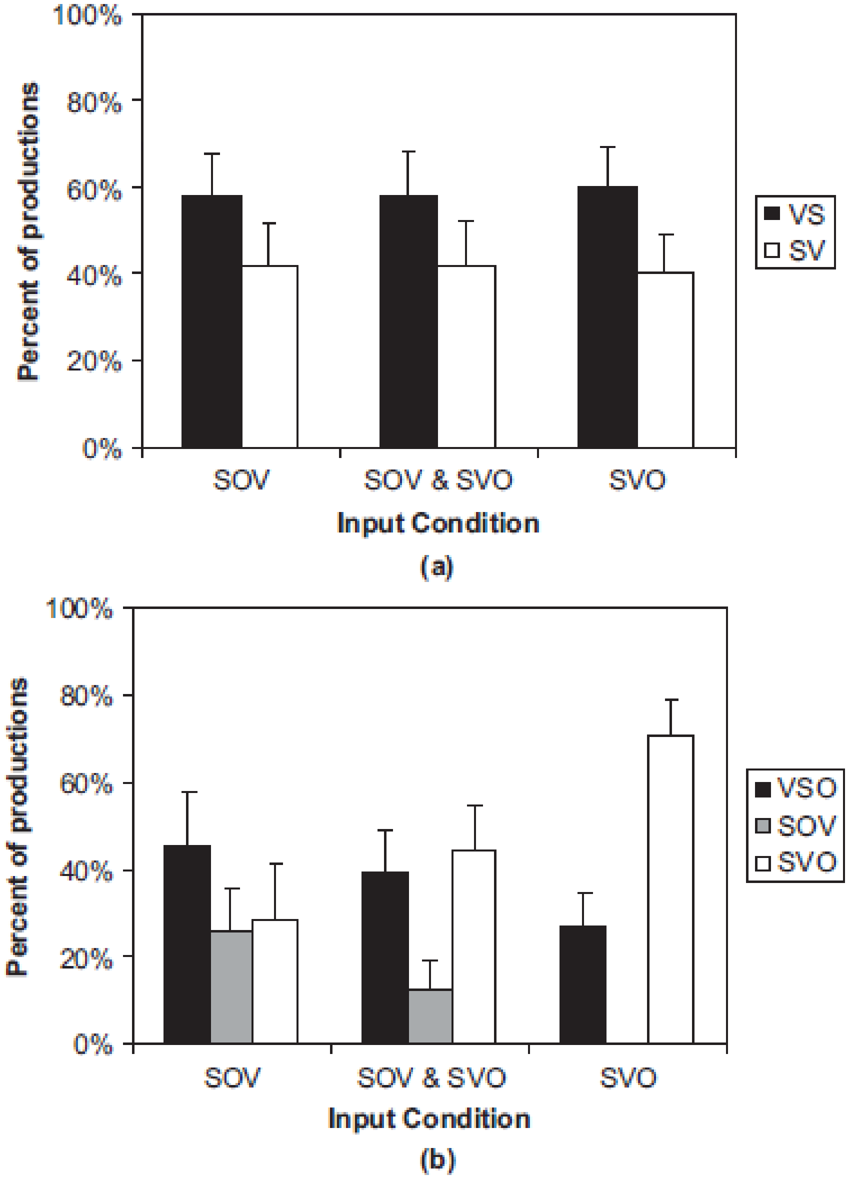
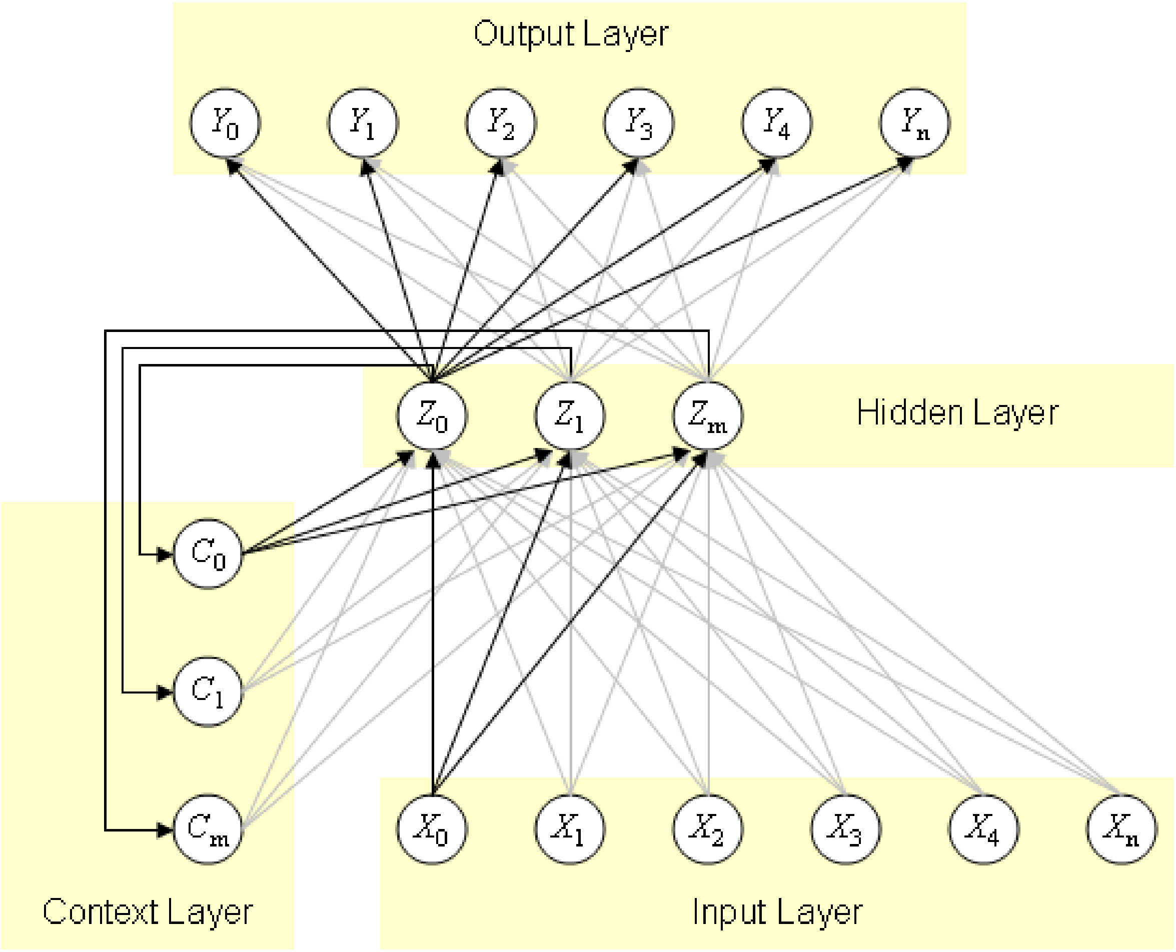


FIGURE 1 Mean production of sentences in different word orders for intransitive (a) and transitive (b) sentences by input condition.

Reproduced from Kam, 2009. The probabilities in (a) are very close to the actual proportions used in the exposure sentences. The results in (b) are not as clean due to L1 interference. Because the subjects were all monolingual English speakers, they were much more likely to use SVO order above the other orders, even when SVO order was never present in the input.

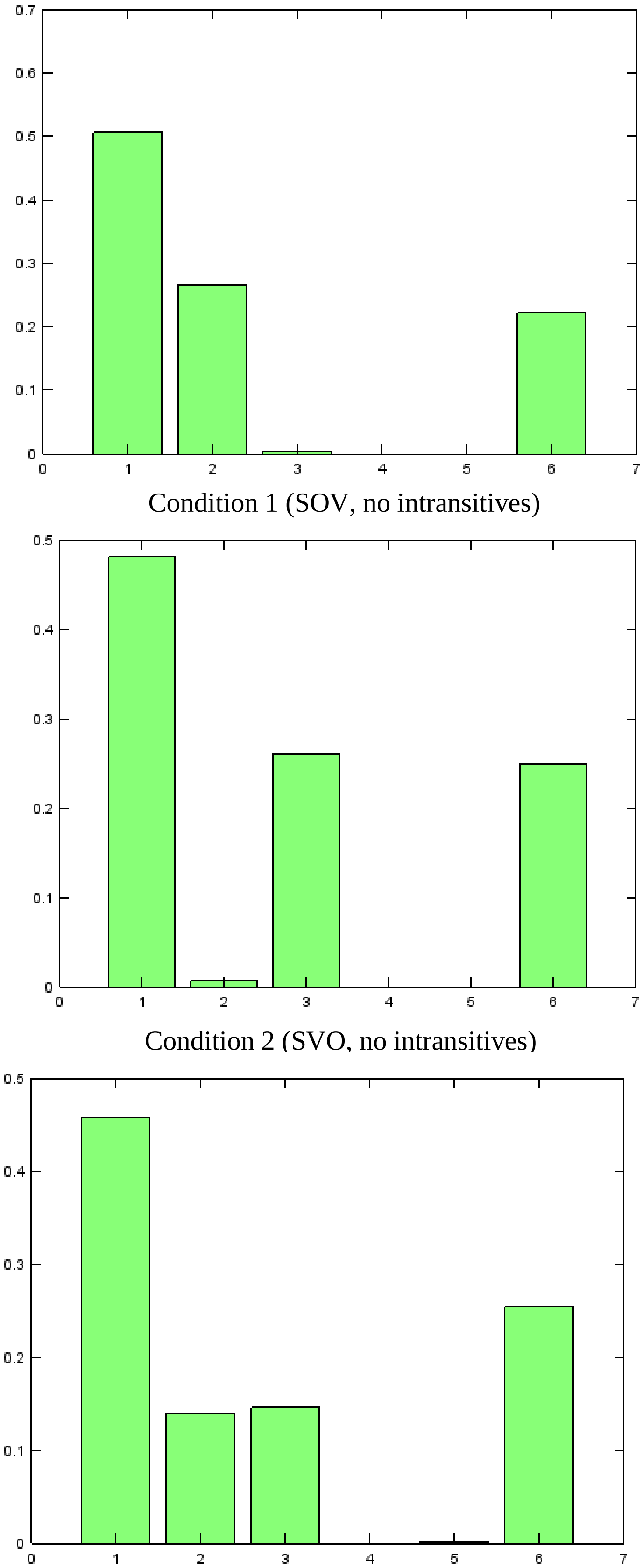
Figure 2



An example of a simple recurrent network. The context layer is set to zero initially; after the first item in the sequence is run, the values from the hidden layer are copied onto the context layer, so that the first input has an effect on the following input. By running successive inputs through the network, sequences can be entrained and recalled later.

Image taken from: http://mnemstudio.org/neural-networks-elman.htm

Figure 3



Condition 1 (SOV, no intransitives)

Condition 2 (SVO, no intransitives)

Condition 3 (SVO & SOV, no intransitives)

Legend:

1: VSO sentences

2: SOV sentences

3: SVO sentences

4: VS sentences

5: SV sentences

6: ungrammatical sentences/errors

These bar graphs show the results from the first 3 conditions. The performance of all three conditions is similar; the grammatical forms follow the distribution inherent in the transitional probabilities from the exposure sentences. By removing the non-grammatical forms from the analysis, the performance can be assessed more accurately.

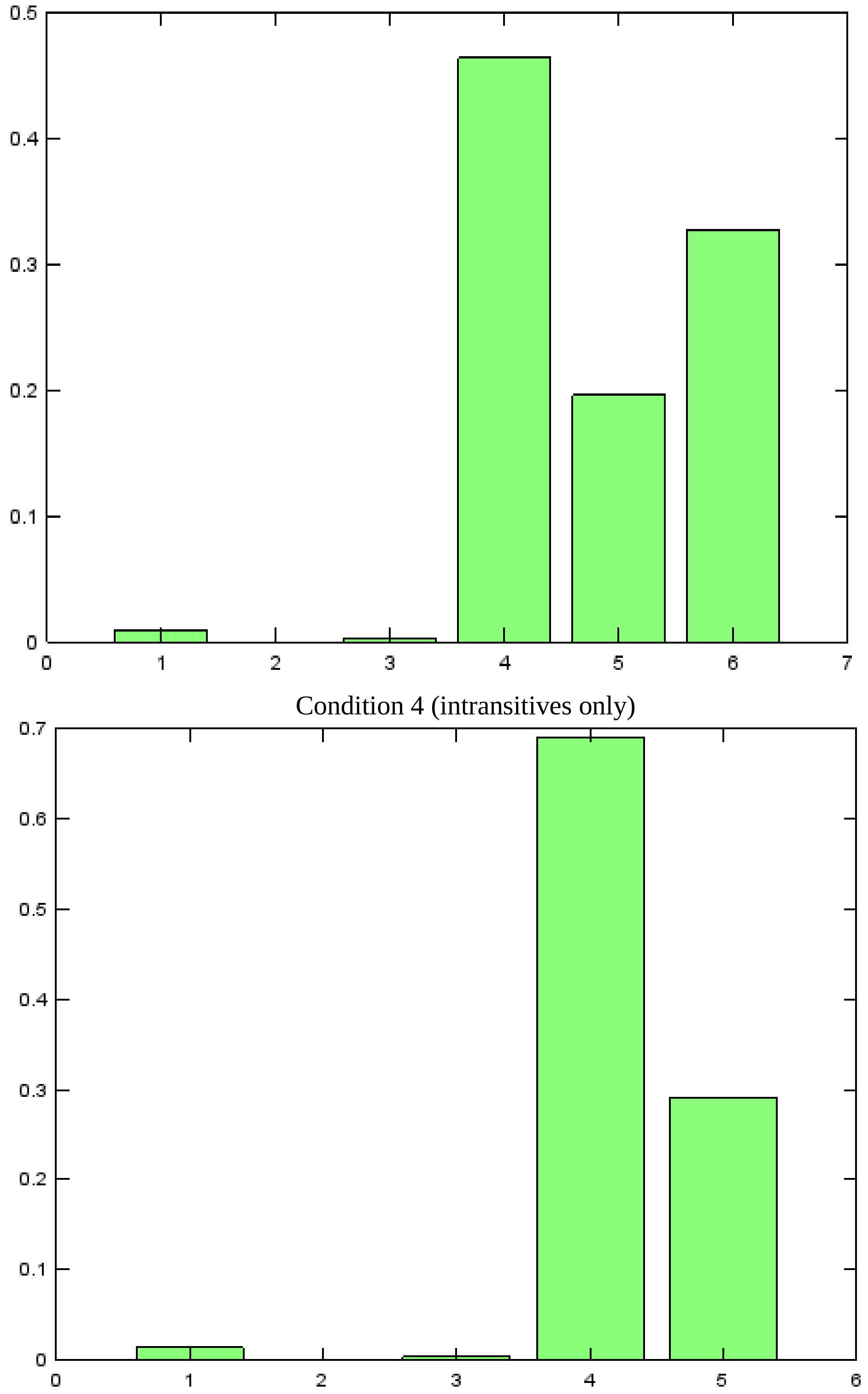
Condition 1: 65%/35%

Condition 2: 64%/35%

Condition 3: 61%/19%/20%

These values match the input probabilities almost perfectly, indicating that the program is actually probability matching, even while still making errors. Though the lack of L1 interference prevents very accurate comparison with the original results (figure 1b), the general trend seems to be similar.

Figure 4

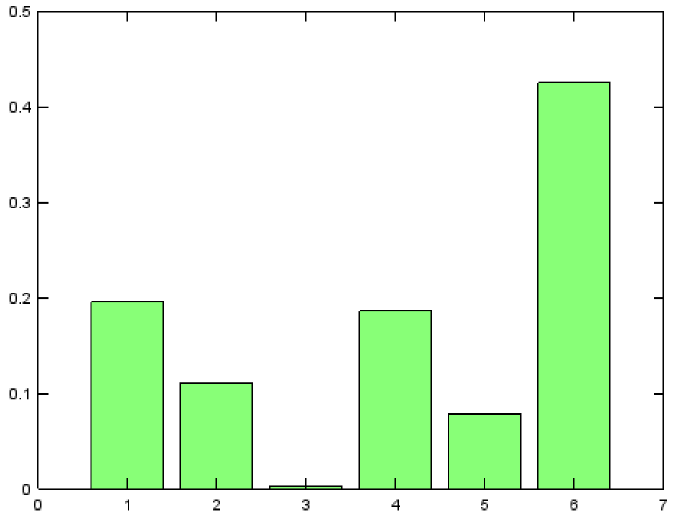


Condition 4 (intransitives only)

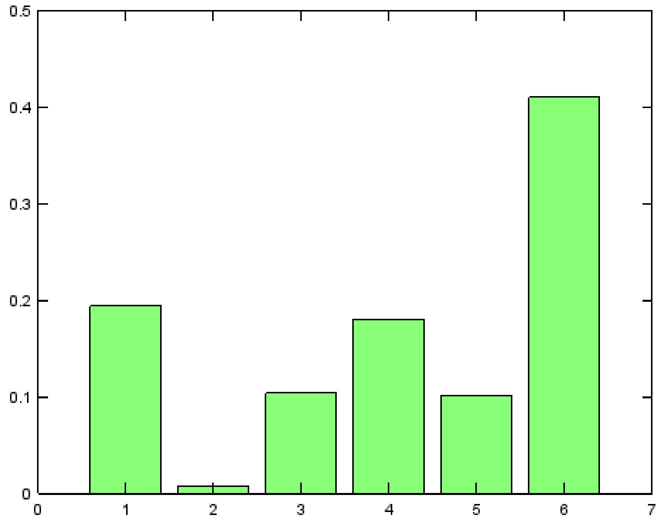
Condition 4 (intransitives only, not including errors)

The probability with which the simulation produces intransitive sentences is a close approximation of the actual proportions of occurrence in the exposure set, mirroring the results of the Hudson Kam experiment (see figure 1a). The second graph shows the results when the errors are not included, demonstrating closer probability matching.

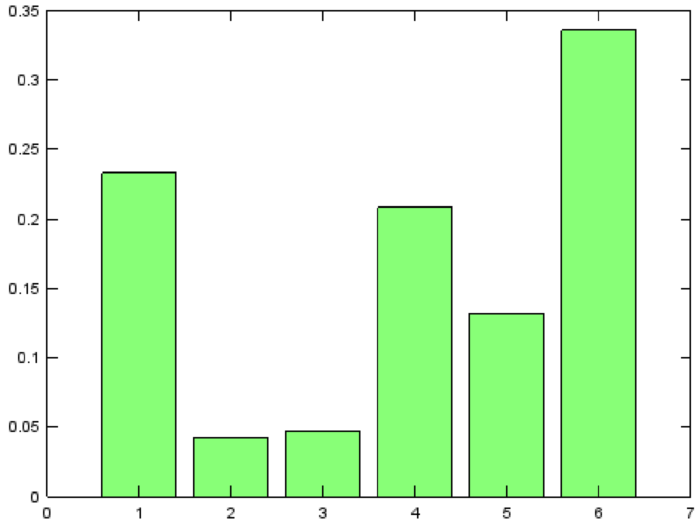
Figure 5



Condition 5 (SOV, with intransitives)



Condition 6 (SVO, with intransitives)



Condition 7 (SVO & SOV, with intransitives)

Legend:

1: VSO sentences

2: SOV sentences

3: SVO sentences

4: VS sentences

5: SV sentences

6: ungrammatical sentences/errors

These graphs show the results for conditions 5-7, which included both transitive and intransitive sentences. They are very similar to the results for conditions 1-3, and match the exposure probabilities fairly well when ungrammatical forms are removed from the analysis.

Condition 5: 70%/38%, 66%/28%

Condition 6: 64%/39%, 62%/39%

Condition 7: 70%/12%/16%, 64%/16%

Though not included in this report, several versions of this simulation were tested. Very clean results can be obtained after very few trials by setting the starting input to a single value for both transitive and intransitive sentences. This approach was scrapped in favor of the current model because the different starting inputs more accurately represent what the output of a previous, linked system would be. Since the model is assumed to be a modular component of a larger system, the differing inputs were included. It should also be noted that the same performance level can be reached when using differing inputs, but many more trials are required to do so.